ISSN 1688-2806

Universidad de la República
Facultad de Ingeniería

Reconocimiento Automático de Configuraciones Manuales propias de las Lenguas de Señas

Tesis presentada a la Facultad de Ingeniería de la Universidad de la República por

Ariel Esteban Stassi Danielli

en cumplimiento parcial de los requerimientos para la obtención del título de Magister en Ingeniería Eléctrica.

Directores de Tesis

Dr. Mauricio Delbracio ..... Universidad de la República
Prof. Gregory Randall ..... Universidad de la República

Tribunal

Dr. Pablo Cancela ..... Universidad de la República
Dr. José Lezama ..... Universidad de la República
Dr. Pablo Musé ..... Universidad de la República
Dr. Marcelo Fiori ..... Universidad de la República

Director Académico

Prof. Gregory Randall ..... Universidad de la República

Montevideo
jueves 4 julio, 2019

Reconocimiento Automático de Configuraciones Manuales propias de las Lenguas de Señas, Ariel Esteban Stassi Danielli.

ISSN 1688-2806

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 172 páginas.
Compilada el jueves 4 julio, 2019.
http://iie.fing.edu.uy/

The deaf believe there is nothing wrong.
The hearing believe something needs to be fixed.

Anonymous

Esta página ha sido intencionalmente dejada en blanco.

Agradecimientos

Quiero agradecer profundamente a Mauricio Delbracio y a Gregory Randall por su paciencia, su confianza y su tiempo, elementos sin los cuales este proyecto no hubiera sido posible.

A Leonardo Peluso, a Ronice Müller de Quadros y al grupo de trabajo de la TUILSU en la Facultad de Humanidades, por acercarme a los fundamentos de la lengua de señas y su documentación.

A Facundo Quiroga, doctorando del III-LIDI, Universidad de la Plata, Argentina, por compartirme su visión sobre el reconocimiento automático de la lenguas de señas y sus principales problemas.

A la familia Álvarez Randall por su calidez humana y, en especial, a Lía Randall por animarme a embarcarme en este proyecto.

A mis compañeros de trabajo de la Licenciatura en Ingeniería Biológica, en especial, a Juan Cardelino por habilitar esta posibilidad, a Germán Pequera por tirar siempre hacia adelante y a Camila Simoes por su ayuda y disposición.

A mi familia, por su acompañamiento incondicional.

Finalmente, y no por ello menos importante, a la Comisión Académica de Posgrado de la Universidad de la República, por el trato recibido y la financiación de este proyecto.

Esta página ha sido intencionalmente dejada en blanco.

A los niños por venir.

Esta página ha sido intencionalmente dejada en blanco.

Resumen

En este proyecto se presenta el estudio de un sistema de reconocimiento de configuraciones manuales propias de distintas lenguas de señas y la evaluación del mismo bajo diferentes condiciones.

En el marco de este proyecto se estudiaron las características fundamentales de las lenguas de señas, esto es, aspectos vinculados a la semántica de una seña como así también a la gramática de este tipo de lenguas. Ello permitió tomar noción de la complejidad propia de este medio de comunicación, y por tanto, de la complejidad ligada a su reconocimiento automático.

La revisión de la bibliografía asociada al Reconocimiento Automático de la Lengua de Señas (RALS) permitió conocer los grandes problemas en este campo, a saber (1) reconocimiento de deletreo manual, (2) reconocimiento de señas aisladas y (3) reconocimiento de discurso continuo, a los cuales se les puede agregar el requerimiento de que el sistema sea independiente del señante. En términos generales, se observó que el RALS es frecuentemente abordado mediante una cadena de procesamiento, compuesta por las siguientes etapas: sensado, preprocesamiento, extracción de características y clasificación. Durante este trabajo se estudiaron distintas variantes para la implementación de cada una de estas etapas, finalizando con la presentación de soluciones basadas en aprendizaje profundo. Dentro de los sistemas más ampliamente utilizados para el reconocimiento de patrones en imágenes aisladas se encuentran las redes neuronales convolucionales (CNN), las cuales se constituyen como redes neuronales de múltiples capas prealimentadas.

La revisión de las bases de datos y las métricas de desempeño permitió tomar noción de los criterios y procedimientos seguidos para la adquisición de un *corpus* con una aplicación particular. Durante esta búsqueda no fue posible encontrar una base de datos de Lengua de Señas Uruguaya (LSU) para el reconocimiento automático. En virtud de ello, durante este trabajo se realizaron dos tareas. Por un lado, se conformó TReLSU-HS, una base de datos para el reconocimiento de configuraciones manuales propias de la LSU a partir de imágenes estáticas. Por otro lado, se sentaron las bases para la adquisición de una base de datos para el reconocimiento de LSU *a nivel de seña*, tomando un subconjunto de Léxico TReLSU como *corpus* de partida.

Durante la etapa de implementación en el marco de esta tesis de maestría se trabajó sobre la reproducción de un sistema de RALS para el reconocimiento de configuraciones manuales a partir de imágenes estáticas. En particular, el sistema *base* utilizado fue Deep Hand, introducido en el artículo “Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and

Weakly Labelled" de Koller y cols., en el cual se implementa y entrena una CNN para el reconocimiento de *configuraciones manuales* propias de la lengua de señas alemana. La metodología seguida para la evaluación de Deep Hand implicó la selección y, eventualmente, la conformación de distintas bases de datos representativas del problema, preprocesadas de acuerdo a los requerimientos de Deep Hand. En particular, se trabajó sobre 4 bases de datos: una base de datos de prueba del sistema Deep Hand; dos bases de datos de deletreo manual, una de lengua de señas alemana y otra de lengua de señas americana; y TReLSU-HS, la cual se introdujo anteriormente.

Sobre las bases de datos listadas, se evaluó el desempeño del sistema Deep Hand, brindando tasas de reconocimiento del orden del 30 % o inferiores. Este hecho motivó la prueba de distintas variantes de *aprendizaje por transferencia*, en las cuales se llevó a cabo el entrenamiento de un clasificador SVM y por K vecinos más cercanos, obteniendo un desempeño del orden del 66 % bajo un esquema independiente del señante, sobre una base de datos de lengua de señas alemana compuesta por 35 clases. Por su parte, las pruebas realizadas sobre TReLSU-HS mostraron un comportamiento fuertemente dependiente de la cantidad de muestras por clase, mostrando la importancia de contar con una base de datos balanceada para la implementación de un sistema para RALS uruguayo de utilidad práctica.

Tabla de contenidos

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 1

Introducción

A lo largo de este trabajo de tesis se realiza una exploración sobre los principales abordajes al problema del reconocimiento automático de lengua de señas. En este capítulo se expone la problemática de este trabajo de tesis y sus posibles soluciones en términos generales. En primer lugar, se introducen las principales características de una lengua de señas, elementos componentes y particularidades de la lengua. En segundo lugar, se presenta el principio de funcionamiento de un sistema típico de reconocimiento automático de lengua de señas. Luego, se expone la motivación para realizar esta tesis y su correspondiente alcance y los objetivos planteados. Por último, hacia el final de este capítulo se lista la organización de este documento.

1.1. Lengua de señas

La *lengua de señas* es el medio de comunicación natural de las personas sordas e hipoacúsicas severas de todo el mundo. De acuerdo con la Organización Mundial de la Salud, se estima que existen 466 millones de sordos en todo el mundo, de los cuales 34 millones son niños [8]. En Estados Unidos, alrededor de 400.000 personas emplean la lengua de señas como primer medio de comunicación [89].

La lengua de señas es un idioma no verbal que emplea una combinación de atributos manuales y no manuales para transmitir un significado y la percepción visual para su recepción [100]. La combinación de estos atributos da lugar a la *seña*, el elemento componente de esta lengua y equivalente a la "palabra" de la lengua oral o escrita [124]. Si bien se emplea en todo el mundo, no se trata de un lenguaje universal, dado que la asociación entre una seña y su significado depende fuertemente de cada región en particular [67].

1.1.1. Características de una seña

Desde el punto de vista lingüístico, la *seña* constituye la unidad mínima aislable de una lengua de señas que porta un significado [124]. En general, una seña se compone de rasgos manuales y rasgos no manuales. A continuación se detalla cada uno de ellos.

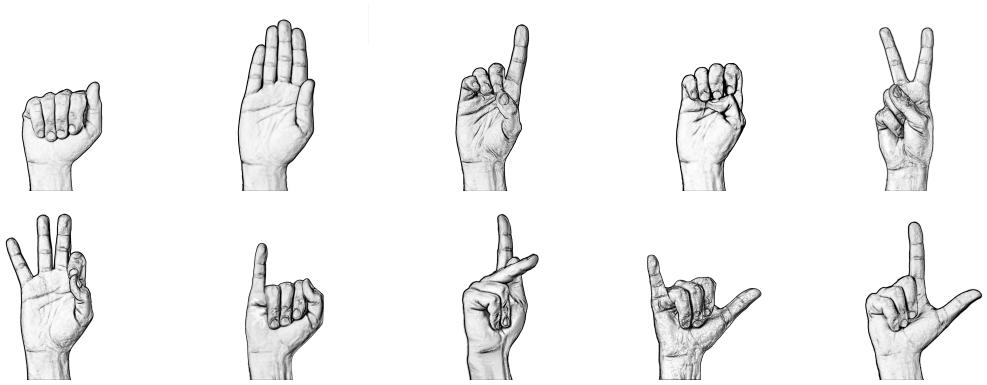


Figura 1.1: Ejemplos de configuraciones manuales. Tomadas de [5].

Rasgos manuales. Comprenden la configuración, la ubicación, la orientación y la trayectoria de una o ambas manos durante la ejecución de una seña [100, 112]:

- La *configuración* [100] –o *handshape* [134]– corresponde a la forma y disposición de los dedos. En general, cada mano puede adoptar configuraciones diferentes y, a su vez, cada configuración puede cambiar conforme se ejecuta la seña. Luego, es frecuente describir una seña a partir de las configuraciones inicial y final adoptadas por cada mano [100]. En la Figura 1.1 se observan algunos ejemplos de configuraciones manuales posibles.

- La *ubicación* [100] –o *hand location* [134]– refiere a la posición o secuencia de posiciones en que se ubican las manos durante la ejecución de la seña. Generalmente, se considera la posición relativa de las manos con respecto a otra parte del cuerpo, tales como los hombros, el abdomen o la cabeza [110].

- La *orientación* [100] –o *hand posture* [134]– alude a la dirección que toma la palma y la punta de los dedos en un determinado momento de la seña. Entre otras, la orientación puede ser "palma hacia arriba", "palma hacia abajo", "palma hacia afuera", "palma hacia adentro" o "palmas enfrentadas" [110].

- La *trayectoria* [100] –o *hand motion* [134]– es el movimiento descripto por cada mano conforme transcurre la seña. Los movimientos que componen la trayectoria suelen ser cortos, precisos y de forma lineal, circular o "zigzagueante" [35, 110]. Al realizar una seña, el señante lleva sus manos desde una posición neutral hacia la posición requerida, mientras configura sus manos según la seña a ejecutar y las señas subsecuentes [75]. Todos los movimientos involucrados en la ejecución de las señas tienen lugar dentro de un espacio virtual llamado *espacio de señado* –o *signing space*– [67].

- Podrían considerarse además otros rasgos manuales, tales como la cantidad de arranques y detenciones de cada mano a lo largo de la seña y los movimientos macro, micro y de contacto, entre otros [55, 100].

En las señas ejecutadas con una sola mano; por ejemplo, durante el *deletreo manual* –o *fingerspelling*–, el señante mostrará una clara preferencia por una de

las manos [131]. En las señas realizadas con las dos manos, dependiendo de la seña, exisitirán rasgos manuales simétricos o asimétricos [55]. En los casos de asimetría es posible identificar una mano dominante y una mano no dominante. La *mano dominante* posee el rol principal o activo dentro de la seña –por analogía, podría decirse que ésta es la mano con la cual el señante "escribe"–, mientras que la mano *no dominante* ejecuta un rol de soporte [55, 131].

Rasgos no manuales. Vienen dados principalmente por expresiones faciales, los patrones labiales, la mirada y las posturas del cuerpo y de la cabeza [134]. Los rasgos no manuales son *esenciales* en la lengua dado que portan información sobre la gramática y la prosodia [134]: para los mismos rasgos manuales, la postura de la cabeza y las expresiones faciales permiten afirmar, negar, preguntar, adherir duda, seriedad, ironía o enojo a la seña que se esté ejecutando [15, 124, 134]. Más aún, en algunos casos los rasgos no manuales permiten resolver ambigüedades en las señas [134]. A modo de ejemplo, tal como se ilustra en la Figura 1.2 en Lengua de Señas Alemana (DGS[1]), 'hermano' y 'hermana' poseen los mismos rasgos manuales y sólo es posible diferenciarlas a partir de sus patrones labiales [134]. Se remarca así la importancia de los rasgos no manuales en la determinación de las señas.

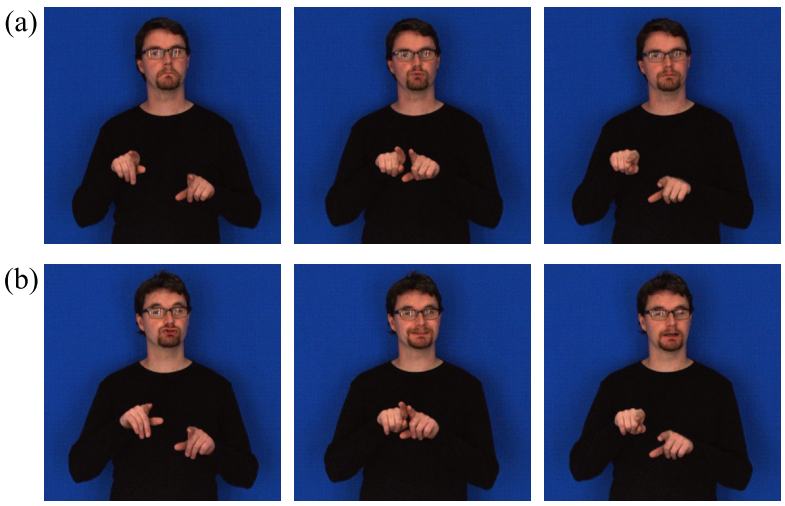


Figura 1.2: En DGS, las señas 'hermano' y 'hermana' son idénticas en rasgos manuales pero difieren en sus patrones labiales. Tomada de [134].

1.1.2. Léxico, Gramática e Iconicidad

Uno de los descubrimientos más importantes del siglo pasado fue que las lenguas de señas usadas por las comunidades sordas son lenguas completas con el mismo poder de expresión que las lenguas habladas, permitiendo comunicar las ideas de forma clara y precisa [124]. Para ello, cada lengua de señas posee los mismos principios

[1]Por sus siglas en alemán, *Deutsche Gebar Gebärdensprache*.

organizacionales que las lenguas habladas [97], entre los cuales se destacan los conceptos de semántica, léxico, gramática e iconicidad.

De acuerdo a la RAE[2], la *semántica* es la "disciplina que estudia el significado de las unidades lingüísticas y de sus combinaciones". En una lengua de señas, cada unidad lingüística se conforma a partir de una combinación única de rasgos manuales y no manuales. La asociación entre su significado y su significante difiere en cada región, dando lugar al concepto de léxico.

De acuerdo a la RAE, el léxico es el "vocabulario o conjunto de las palabras de un idioma, o de las que pertenecen al uso de una región, a una actividad determinada, a un campo semántico dado". En el marco de este trabajo, podría afirmarse que, dada una región, el *léxico* de una lengua de señas queda conformado por un conjunto discreto de asociaciones entre un significado y su correspondiente seña [85].

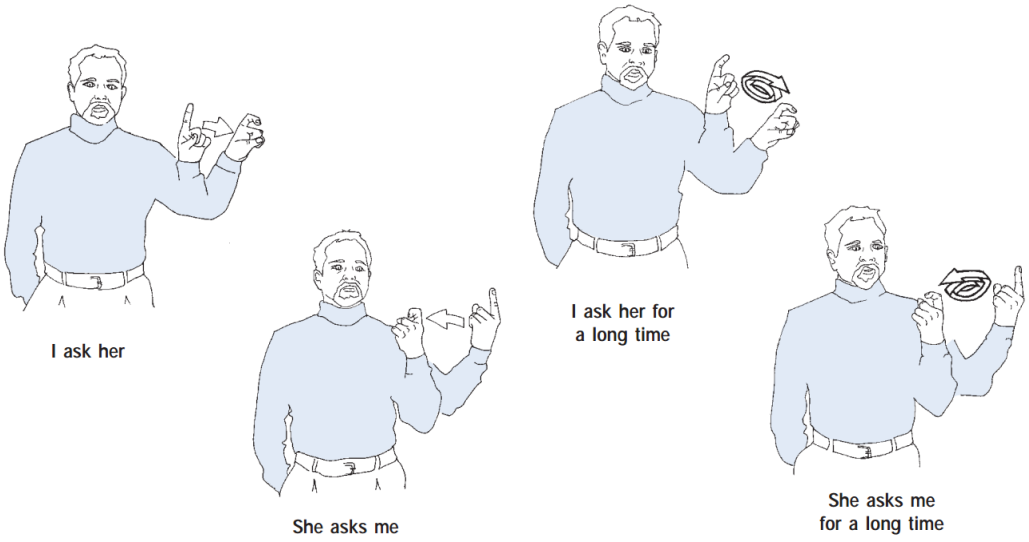
De acuerdo a la RAE, la grámatica es la "parte de la lingüística que estudia los elementos de una lengua, así como la forma en que éstos se organizan y se combinan". La *gramática* de la lengua de señas es radicalmente diferente de la gramática de la lengua oral o escrita. La estructura de un *enunciado* –o *utterance*– oral o escrito es secuencial, es decir, cada palabra, fonema, silencio o elemento componente –en un sentido más amplio– sucede al anterior inmediato. En cambio, un enunciado en lengua de señas posee tanto una estructura secuencial de señas, como una estructura simultánea de los rasgos o elementos componentes de cada una de éstas. La interpretación de cada seña requiere la percepción del comportamiento de ambas manos, en general diferente, sumado a los rasgos no manuales y a la relación entre el sujeto y el objeto [67].

En la Figura 1.3 se observa un ejemplo de construcción gramatical en Lengua de Señas Americana (ASL[3]) sobre la seña 'ask' –en castellano, 'preguntar', 'pedir'–. Observar la configuración manual componente. Por un lado, en la Figura 1.3a se observa un cambio del sujeto y del objeto de la acción mediante un cambio en la orientación y dirección del movimiento de la mano [101]. Por otro lado, en la Figura 1.3b se ilustra una manera de modificar el verbo, en este caso a partir de un complemento circunstancial de tiempo, mediante el agregado de una trayectoria circular a la seña [101].

Por último cabe destacar el concepto de iconicidad, de suma importancia en la construcción –semántica– y el funcionamiento –gramática– de las lenguas de señas. Debido a la percepción visual de la lengua, la *iconicidad* puede definirse como la similitud morfológica entre una seña y su correspondiente referente en el mundo [97]. Luego, una seña puede representar aspectos físicos del referente, su ubicación tridimensional en el espacio, patrones de movimiento y/o referencias temporales [97]. A modo de ejemplo, en Lengua de Señas Francesa, el objeto 'casa' se representa describiendo con las manos la forma de un techo [85]. Se estima que al menos un tercio de las señas de un léxico son icónicas, y que entre el 50 y el 60 % de la estructura de las señas puede ser directamente vinculada a las características físicas del referente [97]. Las señas icónicas y arbitarias existen

[2]Por sus siglas en español, *Real Academia Española*.

[3]Por sus siglas en inglés, *American Sign Language*.



(a) Sujeto y objeto. (b) Modificadores del verbo.

Figura 1.3: Grámatica de la seña 'ask' en ASL. Tomada de [101].

en los léxicos de todas las lenguas [55]. Si bien existen diferencias regionales, es esperable que las señas fundadas en la iconicidad sean similares e incluso iguales entre las distintas lenguas [97]. En la Figura 1.4 se observan cuatro señas de la Lengua de Señas Británica (BSL[4]), ordenadas de izquierda a derecha según su grado de iconicidad: (**A**) transparente, (**B**) translúcida, (**C**) oscura y (**D**) opaca. Por último, es interesante remarcar que el grado de iconicidad es en sí mismo un continuo, ya que representa la facilidad de acceso al significado para un sujeto no señante [97].

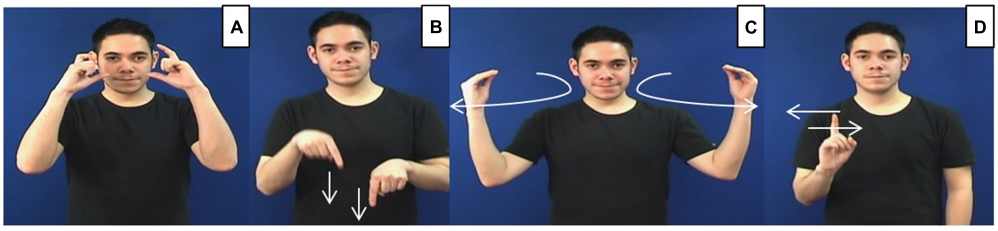


Figura 1.4: Cuatro señas de BSL clasificadas según el grado de iconicidad o *meaning transparency*. **A**: 'cámara'; **B**: 'renguear'; **C**: 'Holanda' y **D**: 'What'. Tomada de [97].

[4]Por sus siglas en inglés, *British Sign Language*.

1.2. Motivación y desafíos

Las personas oyentes frecuentemente no comprenden –o directamente desconocen– la lengua de señas. Luego, para comunicarse con ellas, las personas sordas usualmente deben utilizar el texto escrito, o bien recurrir a un sujeto intérprete intermediario [116]. Si bien es ampliamente utilizada en la comunidad sorda, la escritura resulta poco práctica como medio de comunicación, particularmente durante la caminata, conversaciones a la distancia o cuando hay más de dos personas involucradas en la conversación [116]. Sumado a ello, algunos sordos poseen grandes dificultades para la lecto-escritura [67], lo cual implica un proceso de comunicación aún más costoso. El "uso" de intérpretes intermediarios está seriamente limitado por la poca disponibilidad de ellos y por acarrear un elevado costo y una pérdida de la independencia y la privacidad [67, 116].

Es posible identificar que la comunicación entre las personas sordas y oyentes es complicada. Este hecho motiva el desarrollo de tecnologías de fácil acceso y bajo costo que permitan reducir la brecha existente entre la comunidad sorda y el resto de la sociedad, puesto que aún existen muy pocos oyentes capaces de comprender esta lengua y la adaptación de los sordos a los oyentes no sucede naturalmente. Con el advenimiento de las tecnologías de registro digital, el *Reconocimiento Automático de la Lengua de Señas* mediante computadoras –en adelante denominado RALS[5]– ha cobrado especial interés para atender a este problema. Más específicamente, mediante el RALS se busca:

- incluir a aquellas personas sordas que lo requieran mediante un sistema de traducción automática desde la lengua de señas a la lengua escrita u oral [33, 49].
- desarrollar tecnologías de *software* para la enseñanza de la lengua de señas, sus rudimentos y las particularidades de cada seña [26].
- atender a las necesidades de la comunidad sorda *per se*, ya sea a partir de herramientas de análisis automático de la lengua de señas [50], contribuyendo a la comprensión y al desarrollo de la lengua en sí misma, como así también técnicas para la indexación automática de videos que permitan un acceso más rápido a la documentación basada en lengua de señas [67].

De manera general, el problema del RALS consiste en "identificar mediante una computadora el contenido lingüístico existente en un dato o una secuencia de datos provenientes de un señante". Si bien el foco de este trabajo es el RALS, la investigación en este campo se encuentra muy ligada al reconocimiento automático de gestos manuales estáticos y dinámicos, con lo cual también fue necesario revisar parte del estado del arte en este campo.

El RALS constituye un problema interdisciplinario complejo *aún abierto* y constituye una de las líneas de investigación más importantes en las interfaces hombre-máquina [68]. Dentro de los tópicos que se entrelazan para atacar este

[5]Frecuentemente referido como SLR, por *Sign Language Recognition*.

problema se pueden mencionar las particularidades de las lenguas de señas desde el punto de vista lingüístico, el modelado matemático, el procesamiento digital de imágenes, la interacción hombre-máquina, la visión artificial, el *aprendizaje automático* –o *machine learning*–, el procesamiento del lenguaje natural y el manejo de bases de datos [67, 111].

Los principales desafíos del RALS se encuentran en el desarrollo de métodos robustos frente a [31, 75]:

- el grado de similitud entre las distintas señas;
- la variabilidad espacial de las señas, en un mismo sujeto o entre distintos sujetos –en configuraciones manuales, movimientos u orientación de las manos, contextura física, rostro y gestos faciales–;
- la variabilidad temporal de la dinámica de las señas;
- la variabilidad del entorno, esto es, cambios de *fondo* –o *background*– y de las condiciones de iluminación del señante;
- la posibilidad de oclusiones entre regiones de interés –por ejemplo, en la captura de un sistema basado en imágenes, una mano queda "tapada" por la otra, o una mano se encuentra ocluyendo el rostro–;
- el tamaño del *corpus* a reconocer.

1.2.1. Léxico TReLSU: caso de aplicación en Uruguay

Atendiendo a la falta de estandarización de la Lengua de Señas Uruguaya –en adelante, LSU–, desde la Tecnicatura Universitaria en Interpretación en LSU se ha desarrollado el Léxico TReLSU, el cual constituye el primer diccionario *monolingüe* de LSU, esto es, un *corpus* compuesto por 315 señas propias de la LSU, las cuales se encuentran, a su vez, descriptas en LSU. Léxico TReLSU se encuentra accesible en el web http://tuilsu.edu.uy/trelsu/.

Dado que se trata de una documentación de tipo diccionario, se hace concreta la necesidad de ordenar e identificar las señas componentes de acuerdo a algún criterio de orden y búsqueda. En el caso de una lengua escrita el criterio empleado para organizar un diccionario es el orden alfabético de las letras conforme se avanza sobre la palabra requerida. De esta manera, es posible entablar una relación biunívoca entre una palabra que quiera buscarse y el orden que ésta posee en un determinado conjunto. Sin embargo, en un diccionario documentado en lengua de señas, el sentido de orden puede resultar más complejo. De manera análoga a la organización de las palabras escritas en un diccionario, es lógico pensar que la organización en este caso puede llevarse a cabo a partir de la identificación y la organización de los elementos que constituyen cada seña.

Tal como se explicó en la Sec. 1.1.1, una seña resulta de una composición única de rasgos manuales y no manuales. Dentro de los rasgos manuales se encuentran la configuración inicial y final, el movimiento, la ubicación y la orientación; mientras que los rasgos no manuales vienen dados principalmente por expresiones faciales,

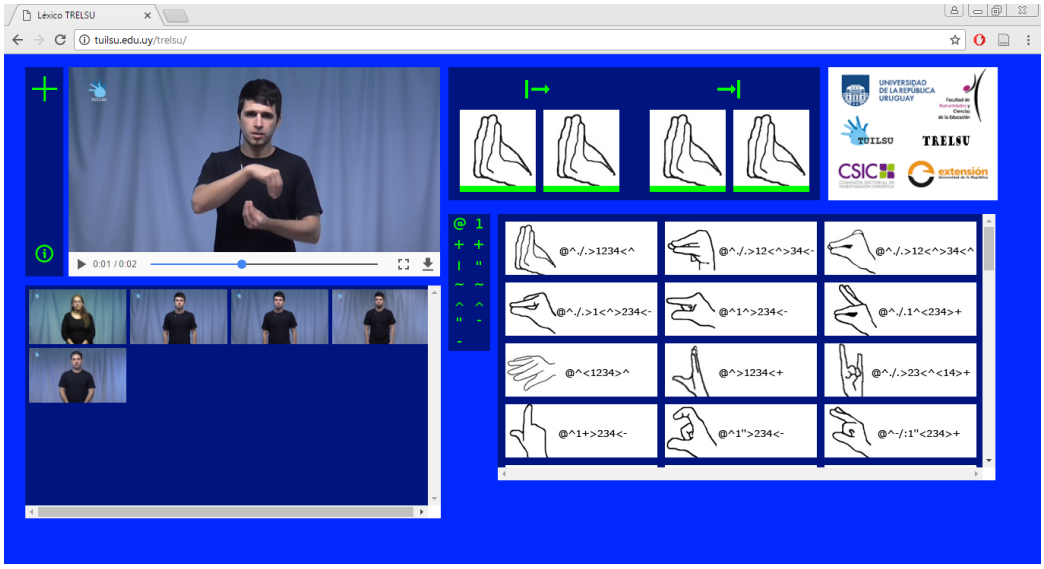


Figura 1.5: Captura de pantalla del sistema Léxico TReLSU. Tomada del sitio *web* http://tuilsu.edu.uy/trelsu/.

las miradas y la postura del cuerpo [100]. Resulta lógico pensar que la consideración de *todos* estos parámetros propios de cada seña redunda en una búsqueda sumamente tediosa. En el caso del Léxico TReLSU, la sistematización de la búsqueda se ha realizado a partir de un código alfanumérico desarrollado *ad hoc* para caracterizar la configuración de cada mano, al comienzo y al final de la seña requerida. De esta manera, la búsqueda se realiza a partir de cuatro entradas por parte del usuario, a saber [100]:

- **configuración inicial, mano dominante:** recuadro inferior-*izquierdo* al símbolo $\boxed{|\rightarrow}$.
- **configuración final, mano dominante:** recuadro inferior-*derecho* al símbolo $\boxed{|\rightarrow}$.
- **configuración inicial, mano no-dominante:** recuadro inferior-*izquierdo* al símbolo $\boxed{\rightarrow|}$.
- **configuración final, mano no-dominante:** recuadro inferior-*derecho* al símbolo $\boxed{\rightarrow|}$.

La introducción de esta información el sistema de búsqueda arroja un subconjunto de señas, entre las cuales se encuentra –idealmente– la seña requerida, tal como puede verse en la Figura 1.5.

Si bien este diccionario es monolingüe, la práctica correcta de los usuarios actuales implica que deben conocer y saber emplear el código –o la iconografía– de búsqueda con el cual se han ordenado las señas. Surgen así algunas preguntas que motivan la realización de este trabajo de tesis:

- ¿es posible mejorar el mecanismo de búsqueda dentro de la base de datos Léxico TReLSU?
- ¿es posible generar un mecanismo de búsqueda más "natural" para el usuario a partir del sensado de la actividad del señante y acoplando al sistema un módulo de reconocimiento automático de las configuraciones inicial y final?
- Más aún, ¿es posible generar un mecanismo de búsqueda de la seña que específicamente interesa, incluyendo no sólo las configuraciones inicial y final de ambas manos sino también el resto de los rasgos característicos expuestos en la Sec. 1.1.1?

1.3. Objetivos

Los objetivos de esta tesis de Maestría en Ingeniería Eléctrica son:

- Profundizar sobre los conocimientos de la lengua de señas y comprender de manera general la problemática de su reconocimiento automático.
- Estudiar las distintas soluciones existentes, abordadas mediante técnicas de aprendizaje automático.
- Reproducir al menos parcialmente un resultado actual sobre el reconocimiento de las configuraciones manuales, siendo éste un problema central del reconocimiento automático de la lengua de señas.

1.4. Comentarios de fin de capítulo

A lo largo de este capítulo se comentaron las principales características de las lenguas de señas, composición de una seña y aspectos asociados a la gramática del lenguaje. Se presentó también una motivación para el RALS y se enunciaron brevemente las dificultades técnicas que deben considerarse en la solución a este problema. En el siguiente capítulo se realizará una exposición sobre los distintos abordajes del RALS, partiendo de una revisión de los principales métodos de captura digital y arribando a los sistemas de clasificación frecuentemente empleados en este campo.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 2

Reconocimiento Automático de la Lengua de Señas (RALS)

En este capítulo se presentan las bases para comprender los distintos abordajes al problema del RALS. Según lo definido en el Capítulo 1, el RALS puede plantearse como la clasificación automática de un dato sensado, el cual puede ser tanto una señal eléctrica unidimensional, una foto o un video que de uno u otro modo miden la actividad del señante.

Dado un dato –o una secuencia de datos– de entrada, el RALS abarca el proceso de segmentación, seguimiento e identificación de las señas ejecutadas y su posterior conversión a palabras o expresiones semánticamente correctas [31]. En la Figura 2.1 se puede observar la *cadena de procesamiento* –o *pipeline*– de un sistema típico de RALS, en el cual se ha supuesto que los datos de salida del sensor ya se encuentran en el dominio digital.

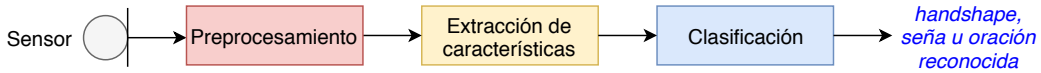


Figura 2.1: Etapas de un sistema típico de RALS. Adaptado de [67].

A continuación se hará una breve descripción de cada una de las etapas involucradas.

- **Sensado.** En esta etapa se adquiere la información digital de entrada al sistema a partir de la cual identificar los rasgos manuales y no manuales de las señas. Mediante este proceso se realizan observaciones del "mundo" a partir de las cuales inferir su comportamiento. Tradicionalmente, este problema fue abordado partiendo de imágenes digitales en RGB, tanto estáticas como dinámicas [67]. En los últimos años, la comunidad científica ha comenzado a incorporar otras técnicas de sensado, tales como medidas basadas en sensores de movimiento y profundidad o el uso de guantes instrumentados con sensores de elongación, orientación y aceleración. En la Sec. 2.2 se describirá cada una de estas técnicas en detalle.

- **Preprocesamiento.** Independientemente de la tecnología de adquisición empleada, en general será necesario segregar sólo la información útil para el reconocimiento. En caso de trabajar sobre gestos estáticos –independientes del tiempo–, esta etapa consiste en la segmentación de las manos y demás partes relevantes del cuerpo para el reconocimiento, tales como el rostro y el torso del señante, entre otros. Por otro lado, en el caso del reconocimiento de gestos dinámicos –dependientes del tiempo– a partir de registros de video, no sólo debe realizarse la segmentación antedicha, sino que además deben aislarse sólo aquellas muestras temporales que aporten la información de interés. Habiendo segmentado los datos, tanto en espacio como en tiempo, es frecuente llevar a cabo una normalización a una disposición espacial y temporal de referencia. Por último, para el caso de gestos dinámicos suele realizarse una etapa de *seguimiento* –o *tracking*– a los fines de considerar cómo se mueve cada porción segmentada a lo largo del tiempo. En la Sec. 2.3 se hará una descripción en detalle de los métodos de preprocesamiento hallados en la bibliografía.

- **Extracción de características.** Mediante este proceso se busca reducir la dimensionalidad de la información de forma eficiente. Para ello se realizan mediciones sobre las imágenes que brinden descriptores numéricos ligados a los aspectos de interés para el reconocimiento. A este proceso se lo denomina "Extracción de características" y en este caso consiste en tomar medidas que describan la geometría de las configuraciones manuales y expresiones faciales, la posición relativa de las manos *versus* un punto anatómico de referencia o la trayectoria trazada por las manos, entre otras. En la Sec. 2.4 se presentarán las características comúnmente empleadas en el abordaje de este problema.

- **Clasificación.** Se trata de un sistema de aprendizaje automático o reconocimiento de patrones que debe ser entrenado para reconocer automáticamente una configuración manual, una seña aislada o una oración en lengua de señas, a partir de características descriptivas provistas como entrada [21]. Lógicamente, el nivel de acierto de esta etapa dependerá fuertemente de la complejidad o nivel de realismo del problema atacado. En la Sec. 2.5 se verán en detalle las distintas configuraciones de los sistemas de clasificación en función del problema a resolver.

A lo largo de este capítulo se hará una descripción de las principales líneas de abordaje de este problema. En particular, se tomará como referencia cada etapa de la Figura 2.1. En la Sec. 2.1 se introducen los principales problemas del área, ordenados según un grado de complejidad creciente. En la Sec. 2.2 se presentan las técnicas de sensado u obtención de datos más empleadas. Luego, en la Sec. 2.3 se comentan distintas estrategias de preprocesamiento, con especial énfasis en la segmentación. Luego, en la Sec. 2.4 se detallan los principales descriptores empleados para abordar este problema. Por último, en la Sec. 2.5 se mencionan las principales técnicas de clasificación del campo.

2.1. Problemas típicos del RALS

El problema del RALS puede ser abordado a distintos niveles de complejidad. En las siguientes secciones presentarán tres casos típicos: reconocimiento de deletreo manual –Sec. 2.1.1–, reconocimiento de seña aislada –Sec. 2.1.2– y reconocimiento de discurso continuo –Sec. 2.1.3–. Por último, en la Sec. 2.1.4 se discute el aspecto "independencia del señante", deseable en general para todo sistema de RALS.

2.1.1. Reconocimiento de deletreo manual

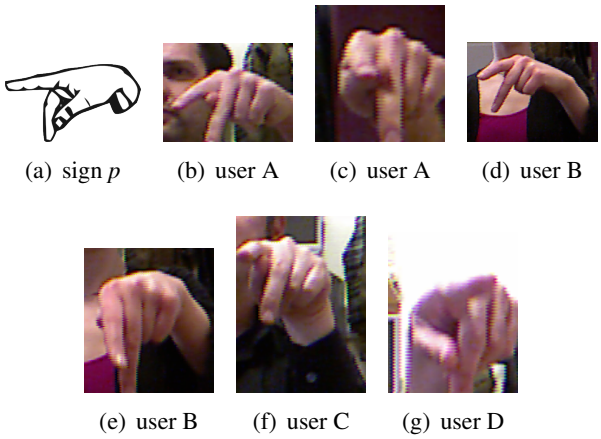
El *deletreo manual* –o *fingerspelling*– de una palabra consiste en la representación manual de cada una de las letras que componen dicha palabra. Para ello, el señante debe hacer uso del alfabeto dactilológico que provee una relación entre ambos dominios de información, al establecer una seña por cada letra del alfabeto escrito u oral [110, 124]. En general, los elementos del alfabeto dactilológico son simples y quedan determinados por la configuración de una sola mano. El alfabeto dactilológico es usado frecuentemente en las lenguas de señas para deletrear nombres, lugares o todo aquéllo que no posea una seña específica [68].

Para el *reconocimiento automático del deletreo manual* –o *fingerspelling recognition*– bastará con el reconocimiento de los rasgos manuales que identifiquen cada elemento del alfabeto dactilológico. En caso de tratarse de alfabetos unimanuales no se tendrá problemas por auto-oclusión [106]. Normalmente cada elemento puede caracterizarse sólo a partir de una configuración manual estática. De esta manera, el reconocimiento de deletreo manual es normalmente abordado a partir de imágenes estáticas. No obstante, algunas letras pueden implicar pequeños movimientos de la mano en su totalidad, motivando el reconocimiento a partir de registros de video. En esta línea surgen, entre otros, el problema de la segmentación temporal de cada una de letras [72]. Se concluye entonces que para el reconocimiento del alfabeto dactilológico se requiere de una descripción precisa de las configuraciones manuales y, en algunas lenguas, del movimiento de las manos [34].

Incluso aunque éste sea el problema más sencillo, puede haber dificultades en cuanto a la variabilidad de las configuraciones manuales entre los distintos sujetos y en cuanto a la gran simulitud existente entre las distintas clases. En las Figuras 2.2 y 2.3 se muestran ejemplos de estos aspectos para el reconocimiento del alfabeto dactilológico de la Lengua de Señas Americana (ASL) [106].

2.1.2. Reconocimiento de señas aisladas

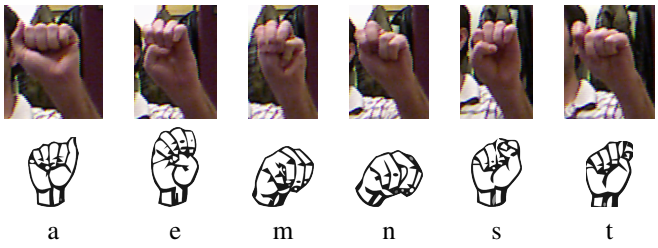
Una *seña aislada* es un elemento del léxico de una lengua de señas, según se definió en la Sec. 1.1.2 de este informe. Cada seña está compuesta por una sucesión particular de rasgos manuales y no manuales. Luego, el reconocimiento automático de señas aisladas en un sistema basado en visión implica la extracción de características de una secuencia de imágenes que capture la dinámica del gesto. En este problema, cada seña se encuentra separada o aislada, ya sea por las condiciones de registro o por haber sido segmentada en tiempo *a priori* [67].



(a) sign *p* (b) user A (c) user A (d) user B

(e) user B (f) user C (g) user D

Figura 2.2: Variabilidad debida a pequeñas variaciones de orientación y diferentes "estilos" de deletreo de la letra 'P' según el señante. Tomada de [106].



a e m n s t

Figura 2.3: Similitud entre las configuraciones manuales correspondientes a las diferentes letras de la ASL. Muestras ejecutadas por el mismo señante. La diferencia entre las clases radica sólo en la posición del pulgar. Tomada de [106].

2.1.3. Reconocimiento de discurso continuo

El procedimiento de clasificación implicado tanto en señas aisladas como en discurso continuo toma en consideración las mismas características. No obstante, el reconocimiento de discurso continuo posee una serie de dificultades adicionales asociadas a la delimitación temporal de las señas. Más específicamente, tanto las oraciones –como sus señas componentes– deben ser secuenciadas de forma automática y, en general, no existe un orden específico para las señas involucradas ni se sabe cuántas señas están contenidas en cada oración [67]. A esto se suma el hecho que en las transiciones entre señas las manos se mueven desde la posición final de la seña ejecutada a la posición inicial de la seña siguiente [34]. Este fenómeno es conocido como *co-articulación* y constituye en sí mismo una línea de investigación tanto para los lingüístas como para la comunidad de las *ciencias de la computación* –o *computer science*–.

Por último, el reconocimiento de discurso continuo implica modelar los aspectos suprasegmentales de la lengua, tales como la prosodia o la gramática. Aún existen muy pocos trabajos que contemplen estos últimos aspectos y quizás se deba a que este problema está abierto y no existe aún una solución satisfactoria [110].

2.1.4. Independencia del señante

Los sistemas de RALS actuales muestran desempeños excelentes operando de manera dependiente del señante, esto es, los datos de prueba del sistema provienen del mismo sujeto señante que los datos con los cuales el sistema fue entrenado durante su implementación [67]. No obstante, en general, la tasa de acierto de un sistema así entrenado decae notablemente cuando los datos de prueba provienen de un señante diferente, esto es, los patrones espaciales y/o temporales presentes en los datos de prueba difieren notablemente de aquéllos presentes en los datos de entrenamiento correspondientes. Este fenómeno se atribuye a la gran variabilidad interpersonal en la ejecución de una lengua de señas. En la Figura 2.4 se pueden observar 5 ejecuciones de la seña 'tenis' en el mismo dialecto de Lengua de Señas Británica (BSL) por parte de 2 señantes nativos. En color negro se ilustra la trayectoria de las manos al realizar cada repetición de la seña.

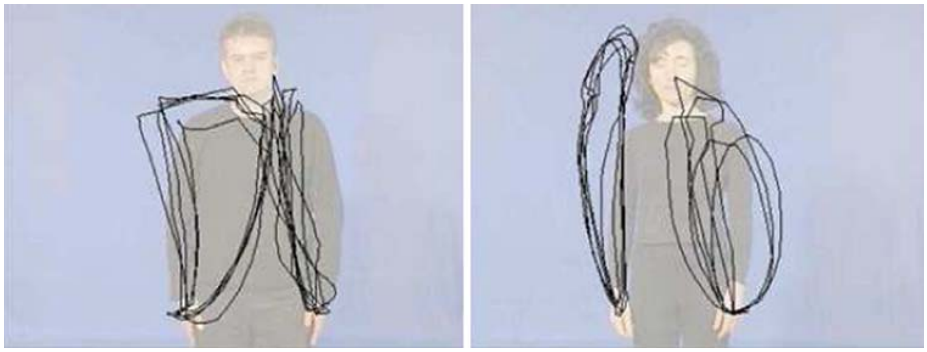


Figura 2.4: Ejecución de la seña 'tenis' en BSL (mismo dialecto), 5 veces por parte de 2 señantes nativos. Tomada de [67].

Si bien la variación entre las distintas realizaciones de la seña por parte de cada sujeto es considerable, resulta mucho mayor la variación existente entre las ejecuciones de ambos sujetos.

El problema de independencia de señante puede reformularse en términos de dos problemas más elementales: el estudio de la variación interpersonal en las señas y la selección de características robustas al señante [34]. En este sentido, se han reportado métodos de adaptación para mejorar el desempeño del sistema de RALS frente a señantes desconocidos [133].

2.2. Técnicas de sensado

El RALS se encuentra estrechamente vinculado al *reconocimiento automático de gestos manuales*[1]. Actualmente, existen diversos dispositivos para realizar el sensado de estos gestos, tales como cámaras RGB convencionales, sensores de profundidad, sofisticados guantes de datos o medidas de acelerometría y electromiografía

[1]Traducido del inglés, *hand gesture recognition*.

sobre los brazos del señante. Dependiendo del tipo de sensor empleado, se contará con diferentes relaciones entre precisión, costo, robustez y practicidad de uso.

En esta sección se expondrán las principales técnicas de sensado utilizadas en este campo, como así también las ventajas y desventajas propias de cada una. Se considerará para ello la taxonomía empleada en el trabajo de Cheok y cols. [31], en el cual los sistemas de reconocimiento de gestos manuales son agrupados en "sistemas basados en visión" y en "sistemas basados en sensores". Por un lado, los *sistemas basados en visión* implican la adquisición de imágenes o videos con la actividad del señante. En este sentido, en la Sec. 2.2.1 se presentarán distintas técnicas de adquisición de imágenes. Por otro lado, los *sistemas basados en sensores* emplean transductores de posición, movimiento y velocidad de los miembros superiores y las manos. En la Sec. 2.2.2 se muestran las técnicas de sensado más difundidas en este tipo de sistemas.

Este trabajo de tesis está centrado principalmente en las técnicas basadas en imágenes. Por esta razón, se desarrollarán con mayor detalle los métodos de RALS basados en visión.

2.2.1. Sistemas basados en visión

Los sistemas basados en visión emplean imágenes o secuencias de imágenes como datos de entrada. Para la adquisición de estas imágenes puede emplearse una o más cámaras, técnicas activas –basadas en el uso de luz estructurada– y técnicas invasivas –basadas en el uso conjunto de cámaras y marcadores sobre el sujeto señante– [31]. En esta sección se describirá cada una de estas técnicas.

Cámaras RGB simples. La técnica más comúnmente empleada en la actualidad para la captura de gestos son cámaras RGB simples, desde sofisticadas cámaras de video a cámaras de teléfonos inteligentes o cámaras *web* de baja resolución [31].

Las cámaras RGB son dispositivos sensibles a la radiación electromagnética en el espectro visible [18]. Para ello, cuentan con un sistema óptico capaz de capturar la luz reflejada sobre un objeto de interés. Luego, la captura requiere de una fuente de luz *externa*. Una vez que la luz ha ingresado a la cámara, es necesario llevar a cabo el proceso de conversión analógica-digital. Para ello, estos dispositivos emplean un arreglo bidimensional de transductores de luz a carga eléctrica en tres canales principales: rojo, verde y azul. Este formato se conoce como RGB y permite codificar los colores del "mundo" mediante un abordaje aditivo de canales, de allí el nombre dado a este tipo de cámaras [59].

Las especificaciones más importantes de una cámara RGB son [18]:

- tipo de elementos del transductor –CCD[2] o CMOS[3]–;
- tamaño de cada elemento en el arreglo del transductor –mayor captura de luz directamente proporcional al tamaño–;

[2]Por sus siglas en inglés, *Charged Coupled Device*.
[3]Por sus siglas en inglés, *Complementary Metal-Oxide-Semiconductor*.

- cantidad de cantidad de columnas y filas del transductor –resolución de píxeles, usualmente medida en *megapixels* (Mp)–;
- profundidad de color –medida en *bits per pixel* (bpp), usualmente 24–.
- distancia focal –medida en *mm*– y apertura –número F– del sistema óptico.

En la Figura 2.5 se puede observar una fotografía RGB de un señante ejecutando un gesto estático.

Figura 2.5: Captura de un gesto estático mediante una cámara RGB convencional.

En caso de capturar gestos dinámicos a partir de video es también importante tener en cuenta la frecuencia de muestreo, comúnmente medida en *cuadros por segundo* –o *frames per second*– y referida como 'fps'. En este sentido, existen valores estandarizados según las distintas normas de captura y transmisión de video, los cuales son 30 fps para NTSC y 25 fps para PAL, de uso masivo en Estados Unidos y Europa, respectivamente [18].

Cámaras térmicas. Las cámaras térmicas conforman la imagen a partir de un perfil bidimensional de temperaturas en la escena. Más específicamente, a partir de la energía calórica radiada por un objeto de interés, proveen información sobre su forma y su posición en la escena. En la Figura 2.6 puede observarse de forma comparada una imagen RGB convencional (izquierda) frente a una imagen capturada por una cámara térmica (derecha). Una gran ventaja de las cámaras térmicas frente a una cámara RGB convencional es su relativa insensibilidad frente a la iluminación ambiente, permitiendo mejorar el desempeño de sistemas de reconocimiento de gestos en situaciones de baja iluminación u oscuridad [18]. No obstante, este tipo de sensores no son muy empleados debido a su elevado costo [18].

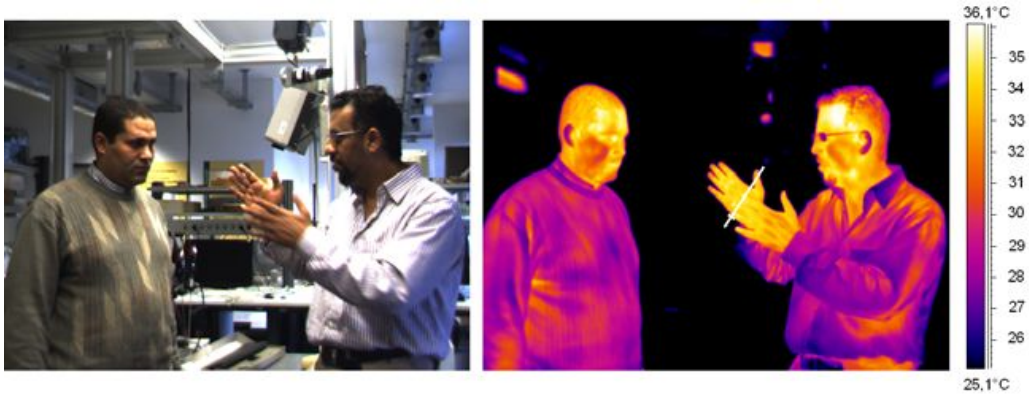


Figura 2.6: Captura mediante una cámara RGB convencional (izquierda) y mediante una cámara térmica (derecha). Tomada de [12].

Si bien una gran parte de los sistemas de reconocimiento se han basado en cámaras 2D, existe una tendencia en el campo de incorporar información de la profundidad de la escena, dando lugar a sistemas de adquisición de imágenes tridimensionales. Entre las limitaciones más importantes de las imágenes 2D para la tarea del RALS se mencionan: (1) las imágenes 2D correspondientes a una seña dependen fuertemente del punto de vista o perspectiva dada por la disposición cámara-sujeto, haciendo más difícil el reconocimiento del gesto ejecutado [31, 87]; (2) en algunos casos la proyección de la seña sobre el plano imagen puede contar con auto-oclusiones –por ejemplo, en la escena una o ambas manos quedan detrás del cuerpo–, hecho que impide "visualizar" la mano en tal caso [87]. Como se verá posteriormente en la Sec. 2.6, este último hecho ha sido la motivación de algunos métodos de reconocimiento basados en el empleo de modelos para inferir la localización de la mano cuando no se encuentra visible desde la perspectiva empleada.

El principio de operación de los sensores ópticos tridimensionales puede basarse en los siguientes mecanismos: visión estéreo (cámaras estereoscópicas), uso de luz estructurada (técnicas activas) y tiempo de vuelo[4] [140]. A continuación se comenta con mayor profundidad el funcionamiento de los dos primeros mecanismos.

Cámaras estereoscópicas. El empleo de 2 o más cámaras permite la obtención de la profundidad de las superficies presentes en la escena [64]. Para ello el arreglo de cámaras debe adoptar una configuración particular y deben conocerse la posición relativa de las cámaras involucradas y sus parámetros internos [64]. Además, las cámaras deben sincronizarse de modo de efectuar capturas simultáneas de la escena. Bajo estas condiciones, es posible obtener la profundidad de la escena mediante algoritmos de *stereo matching* [115] para 2 cámaras; o *bundle adjustment*, para más de 2 cámaras [64]. En la Figura 2.7 se observa un ejemplo de una cámara estereoscópica comercial actual de alta velocidad.

[4]Traducido del inglés, *time-of-flight*.

Figura 2.7: Cámara estereoscópica Horseman 3D Stereo. Imagen tomada de https://www.fotocasion.es/.

Técnicas activas. En contraste con las técnicas pasivas[5], existe un conjunto de técnicas activas que, a partir del uso de luz estructurada, capturan el perfil 3D de la escena. Más precisamente, el perfil se estima a partir del análisis de la "deformación" local que sufre un patrón lumínico conocido sobre las superficies de la escena [140]. Tal es el caso de las plataformas Microsoft® Kinect™, Leap Motion e Intel® RealSense™. A continuación se describen los dos primeros, no así Intel® RealSense™ por su gran similitud con Microsoft® Kinect™. Además de la información de profundidad, algunas de estas plataformas suelen entregar una imagen RGB, con lo cual son frecuentemente referidas como "cámaras RGB-D[6]".

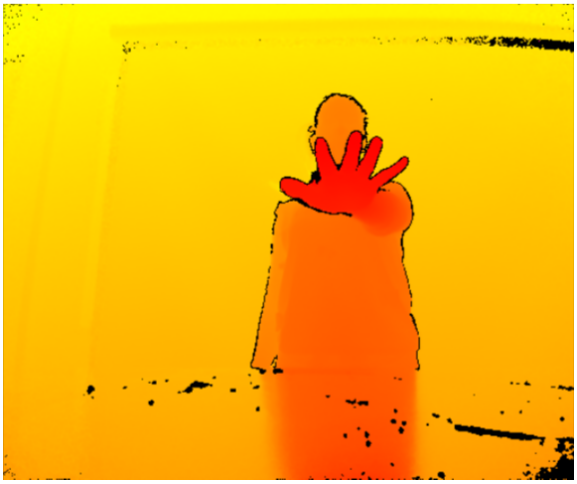
Microsoft® Kinect™ se compone de un proyector de luz infrarroja, una cámara RGB y una cámara infrarroja. La profundidad se estima por triangulación, empleando una matriz de puntos de luz infrarroja proyectada sobre la escena y las imágenes captadas por ambas cámaras [148]. En la Figura 2.8a puede observarse el aspecto general del dispositivo. Asimismo, en las Figuras 2.8b y 2.8c se muestran, respectivamente, el *mapa de profundidad* –o *depth map*– y la reconstrucción 3D correspondiente a la Figura 2.5. En particular, en la Figura 2.8b se hace uso de una paleta de colores que varía desde el rojo hasta el amarillo, conforme aumenta la profundidad de la escena. Entre sus ventajas más importantes pueden mencionarse que a partir de un modelo de segmentos articulados [141] en 20 nodos, este dispositivo captura muy bien el comportamiento del cuerpo, permitiendo reconstruir trayectorias tridimensionales de los gestos corporales dinámicos, incluso bajo condiciones de iluminación variable o fondos complejos [102, 148].

Una limitación seria de este dispositivo en el marco del RALS es su baja resolución para capturar los gestos manuales con detalle. Este hecho motiva su uso conjunto con dispositivos más específicos tales como Leap Motion, el cual fue especialmente diseñado para la identificación de posturas y gestos manuales [53,87].

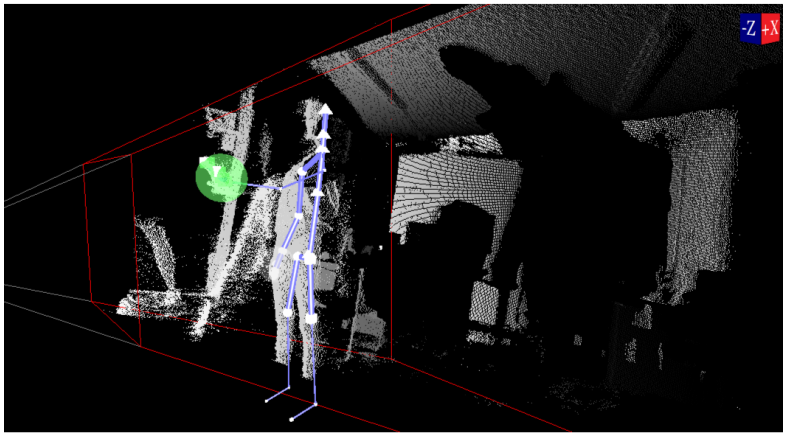
[5]Técnicas de formación de la imagen que no requieren de una fuente de energía lumínica interna.

[6]Por sus siglas en inglés, *Red, Green, Blue and Depth*.

(a) Aspecto general del dispositivo de captura.



(b) Mapa de profundidad del sujeto señante de la Figura 2.5.



(c) Reconstrucción 3D del esqueleto del sujeto señante de la Figura 2.5.

Figura 2.8: Microsoft® Kinect™ y salidas correspondientes.

Por su parte, Leap Motion es un sensor especialmente diseñado para capturar la posición y el movimiento tridimensional de ambas manos y dedos [140]. Leap Motion está compuesto por dos cámaras monocromáticas y tres diodos emisores de luz infrarroja. Además, sobre la información sensada ajusta un modelo para proveer estimaciones de la posición de ambos antebrazos, manos y dedos. Para ello, dicho modelo hace uso de proporciones anatómicas y observaciones pasadas, lo cual permite buenas estimaciones incluso cuando no todas las partes se hallan

Figura 2.9: Sensor de gestos manuales Leap Motion. Tomada de [2].

visibles [3]. Leap Motion cuenta con precisión sub-milimétrica, un ángulo de visión de 150º y provee 28 características por mano, incluyendo posición y orientación tridimensional de la palma de la mano, y la posición y orientación tridimensional de las articulaciones y falanges de cada uno de los dedos [3, 31, 140]. Desde su introducción en el mercado ha llamado especial atención en la comunidad científica para el problema del RALS [86, 87, 91, 104]. En la Figura 2.9 se puede observar un esquema del dispositivo de captura y de la detección de la postura manual que el mismo brinda.

Técnicas invasivas. Existen una serie de técnicas basadas en el empleo de marcadores corporales para la correcta identificación de las regiones de interés y posterior reconocimiento, tales como guantes, pulseras o luces de colores predefinidos [31]. Como se verá en el Capítulo 3, es frecuente el uso de este tipo de marcadores durante el registro de base de datos para el RALS. En general, se emplean guantes de color uniforme, con un tono diferente para cada mano [58, 111, 112, 122]. A los fines de ilustrar las ventajas de emplear este tipo de marcadores, en la Figura 2.10 se muestra un guante con un patrón de *patches* de distintos colores, que no sólo permite la detección de la mano en la escena sino también la desambiguación entre el dorso y la palma –ver la asimetría en los patrones de *patches* según lado– [138].

Figura 2.10: Guante de color diseñado específicamente para el reconocimiento de la postura manual. Tomada de [138].

En [23] los autores proponen un sistema portátil para el RALS basado en una computadora vestible, acelerómetros y el uso de una cámara RGB montada en la visera de una gorra, la cual brinda imágenes en "primera persona"[7]. Para la identificación de las manos, emplean dos pulseras, una por cada muñeca, tal como puede observarse en la Figura 2.11.

(a) Montaje de cámara. (b) Ejemplo de captura.

Figura 2.11: Sistema portátil de RALS propuesto en [23].

2.2.2. Sistemas basados en sensores

La gran variedad de gestos manuales posibles, la variabilidad anatómica de los señantes, las auto-oclusiones de regiones de interés o la adquisición de imágenes a escalas diferentes son algunos de los desafíos a los cuales deben enfrentarse los sistemas basados en visión [14]. Desde los orígenes del RALS han existido soluciones basadas en dispositivos que "por contacto" brinden información precisa sobre la actividad del señante [14, 52, 82]. Sea de forma independiente o como complemento de un sistema basado en visión, los *sistemas basados en sensores* emplean

[7]La etapa de la cámara fue originalmente propuesta por Starner y cols. en [121].

transductores de posición, movimiento y velocidad de los miembros superiores y las manos. Los más comunes dentro de esta categoría son los guantes de datos, aunque también se han encontrado el uso de sistemas basados en acelerometría y electromiografía de antebrazo. A continuación se comenta con mayor detalle en qué consite cada una de estas técnicas de sensado.

Guantes de datos. También llamados guantes activos [105], los *guantes de datos* –o *data gloves*– son plataformas de sensado multicanal especialmente diseñadas para medir la dinámica de la mano a dos niveles. Por un lado, permiten medir la orientación y la aceleración lineal y angular de la mano de forma global mediante un acelerómetro y un giroscopio, denominados conjuntamente IMU[8] [31, 116]. Por otro lado, permiten medir la actividad de la palma y de cada dedo de forma independiente a partir de un arreglo de galgas extensiométricas o de elementos piezorresistivos como elemento sensor [18].

En la Figura 2.12 se muestra un ejemplo de guante comercial, denominado 5DT Data Glove Ultra Wireless Kit, el cual permite interactuar con una computadora via Bluetooth a una distancia máxima de $20\,m$ y cuenta con una autonomía de 8 horas. Dependiendo de la precisión requerida, existen modelos con 1 o 2 sensores por dedo. En [105] se presenta el desarrollo histórico del reconocimiento de gestos manuales, donde es posible observar las distintas generaciones tecnológicas de los guantes de datos.

Figura 2.12: Guante de datos comercial. Tomada de http://www.5dt.com/data-gloves/.

[8]Por sus siglas en inglés, *Inertial Measurement Unit*.

Los guantes de datos proveen información precisa sobre la dinámica de las manos y, en particular, permiten superar el problema de la auto-oclusión comentado previamente en la Sec. 2.2.1 [14]. No obstante, el uso de los guantes posee ciertas desventajas. En primer lugar, limita físicamente los gestos del señante. En segundo lugar, se trata de sensores de alto costo. En tercer lugar, no brindan información sobre la posición de las manos con respecto al cuerpo del sujeto. Por esta razón este tipo de sensores se emplean a menudo de forma conjunta con alguna técnica de adquisición de imágenes como una cámara RGB o RGB-D tal como Microsoft® Kinect™ [14, 49, 86, 87]. Por más información sobre sistemas basados en guantes de datos, dirigirse a [40].

En relación al alto costo de los guantes de datos, se han propuesto alternativas de guantes para sistemas basados en visión que, a su vez, conserven las bondades propias de los guantes de datos, tales como la observación independiente de la actividad de cada dedo. Un ejemplo de este caso es el prototipo de IBDG[9] que se muestra en la Figura 2.13 [98].

Figura 2.13: Prototipo de IBDG: dedales y cámara montada en muñeca. Tomada de [98].

Electromiografía y Acelerometría. Existen algunos trabajos que buscan medir el comportamiento de los miembros superiores del señante mediante el registro simultáneo de la electromiografía y la acelerometría [23, 116, 147]. Por un lado, la *electromiografía* es el registro de la actividad eléctrica de los músculos involucrados en cierta tarea, mediante el uso electrodos de superficie en contacto eléctrico con la piel [24]. Por otro lado, la *acelerometría* es la medición de la aceleración de un objeto en el espacio. En particular, para el caso de RALS se busca medir la actividad de la musculatura que controla la posición y orientación de las manos y dedos [116].

Otras. En la bibliografía se reportó el uso de sensores acústicos basados en ultrasonido y el efecto Doppler, como así también el uso de radares [18]. Ha quedado fuera de los alcances de esta tesis la descripción de estas técnicas.

[9]Por sus siglas en inglés, *Image Based Data Glove.*

2.3. Técnicas de preprocesamiento

Tal como se vio en la Sec. 2.2.1, en los sistemas basados en visión, la información de entrada a esta etapa será una imagen o una secuencia de imágenes. El objetivo de la etapa de preprocesamiento en un sistema de RALS consiste en mejorar el desempeño global en el reconocimiento [31]. En este sentido, es posible identificar distintos tipos de preprocesamiento tales como filtrado, segmentación y normalización [31]. A continuación se describen distintas variantes para llevar a cabo el proceso de segmentación, siendo éste el tratamiento de los datos de mayor relavancia en el marco de este trabajo.

2.3.1. Segmentación

La *segmentación* es el proceso mediante el cual se subdivide una escena en una o más regiones u objetos de interés [59]. En el marco de un sistema de RALS esta técnica se emplea para "alimentar" las etapas subsecuentes sólo con la información relevante, como podrían ser las manos y el rostro, idealmente aisladas mediante la segmentación. Siempre y cuando sea posible una buena segmentación, esta etapa brinda además una mayor robustez del sistema de RALS frente a los cambios de fondo del señante.

En un sistema de RALS la estrategia de segmentación estará fuertemente ligada a la técnica de sensado y las condiciones de adquisición de los datos de entrada. En el caso de capturar la actividad en 3D de un señante, es posible llevar a cabo la segmentación de las regiones de interés según el mapa de profundidad. En [62] se lleva a cabo la segmentación de las manos a partir de un sistema de visión estéreo, tomando como hipótesis que la mano brinda el punto más cercano al elemento sensor. En el caso de capturar la actividad de un señante a partir de una imagen color 2D, será conveniente llevar a cabo una segmentación en el espacio de color [59].

La segmentación en el espacio de color consiste en separar una región de interés en la escena –en este caso, los guantes del señante– a partir de la posición en el espacio de color de los píxeles de dicha región [59]. En este sentido, se han reportado métodos de segmentación no sólo en el espacio RGB sino también en los espacios HSV, HSI y YCbCr, entre otros [28, 31, 63, 103, 108, 118]. Este hecho requerirá la conversión de los datos de un espacio de color a otro como preprocesamiento.

En general, el método de segmentación por color se implementa en tres etapas. En primer lugar, se requiere una caracterización de las regiones de interés en el espacio de representación de color a partir de la cual se establece una serie de umbrales. Es deseable que la determinación de estos umbrales tome en consideración la variabilidad de color de las regiones de interés. En segundo lugar, a partir de los umbrales se lleva a cabo la binarización de la imagen que conforma las máscaras para segmentar. En tercer lugar, se aplican las máscaras sobre la imagen de entrada. Existe también la posibilidad de reemplazar la primera de las tres etapas por un ajuste manual de los umbrales para cada imagen, tal como se reporta en [30]. No obstante, se considera que este abordaje no es compatible con el diseño de un sistema de RALS en tiempo real.

La *detección del color de piel* –o *skin color detection*– es un caso particular de la segmentación por color y consiste en la separación de los píxeles de piel de los píxeles de "no-piel" [118]. Se trata de un problema sumamente complejo debido a la gran variabilidad del color de la piel humana, incluso para una misma persona de una región a otra o bajo diferentes condiciones de iluminación [118]. Este hecho ha motivado el uso de guantes o pulseras para llevar a cabo el proceso de segmentación de un modo más controlado. Este es el caso de los trabajos de [23,58,111,112], tal como se vio en el apartado de técnicas invasivas de la Sec. 2.2.1. No obstante, el uso de guantes o pulseras implica restricciones en el movimiento de las manos y los dedos del señante, lo cual podría dificultar la correcta ejecución de las señas.

Chen y cols. [28] llevaron a cabo la segmentación de manos sin guantes en el espacio RGB. Para ello, detectan las regiones con piel en dos etapas. En primer lugar, obtuvieron una imagen binaria por segmentación en el espacio RGB, a partir de la condición R>G>B, la cual permite aislar las regiones de color rojo, rosa, marrón y naranja. En segundo lugar, realizaron un posprocesamiento de la imagen binaria para conservar sólo aquellos *blobs* similares a un color de piel almacenado previamente. Además de la información de color, el método toma en cuenta la posición y el movimiento de la mano según detecciones pasadas y los bordes de la escena para discriminar la mano del antebrazo.

En [30] presentan una técnica denominada *segmentación de dedos* –o *finger segmentation*–. Los datos de entrada son imágenes RGB tomadas bajo condiciones de iluminación y fondo constantes. Luego, el fondo de estas imágenes es idéntico. Bajo estas hipótesis, es posible segmentar la mano a partir de un método de sustracción de la imagen actual y el fondo. Luego, los autores llevan a cabo un posprocesamiento de la detección mediante segmentación de la piel por color, similar al descripto previamente. Sobre la mano segmentada, estiman la región correspondiente a la palma y, a partir de ésta, los dedos.

En [121] los autores utilizan un modelo de segmentación por color de la piel, en un sistema basado en visión en primera persona similar al presentado en la Figura 2.11. La disposición de la cámara permite capturar la nariz del sujeto de forma fija abajo al centro. Luego, es posible usar el color de este *blob* como un dato para calibrar el modelo de segmentación por color de piel de las manos y dotar al sistema de cierta robustez frente a los cambios de iluminación.

En [16] los autores segmentan la mano a partir de un abordaje de detección del color de piel, combinando la representación del color en los espacios RGB, HSV y YIQ en un vector de características de entrada a un clasificador de tipo perceptrón multicapa. En [36] se propone segmentar las manos en videos *via* detección del color de piel, bajo la hipótesis de que en todo *frame* existe una región con el tono de la piel y que no hay otros objetos en la escena con dicho tono.

En casos de fondo variable en registros de video, puede utilizarse una medida basada en la mediana de varios *frames* para la confección de un modelo del fondo. Luego, mediante el uso de un umbral y midiendo las diferencias de cada píxel de la imagen de entrada con respecto a este modelo, es posible obtener una máscara para aislar el sujeto del resto de la escena [67]. Pisharady y cols. [103] proponen un método para la segmentación de posturas manuales bajo fondos complejos y

naturales basado en un modelo bayesiano de atención visual, alimentado con el color, la forma y la textura de las imágenes. Dicho modelo permite obtener un mapa de probabilidades *a posteriori* denominado *mapa de saliencia*[10], a partir del cual se realiza la segmentación. En la Figura 2.14 se muestra una serie de mapas de saliencia a partir de los cuales segmentaron la mano bajo distintos fondos. Puede observarse en este último caso que el método funciona muy bien, incluso bajo fondos donde ocurre el mismo color que posee la piel.

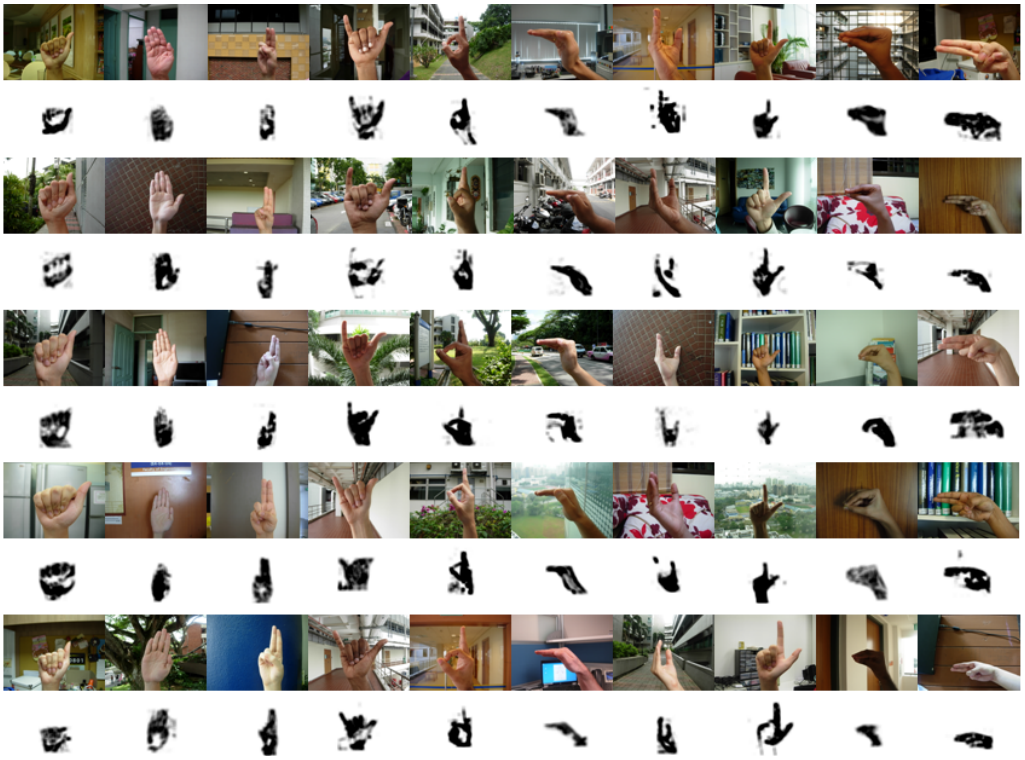


Figura 2.14: Imágenes de muestra y su correspondiente mapa de saliencia. Este último se emplea para la segmentación de la mano. Tomada de [103].

En cuanto al espacio de color empleado, en general se prefiere el espacio HSV frente al RGB, dado que el canal de *tono* –o *hue*– H empleado de forma individual permite representar los colores de forma conveniente [30, 38, 58, 59, 63]. Además, el canal de saturación S puede emplearse para aislar determinados niveles de los tonos de interés en el canal H [30]. La variación de la iluminación de la escena, implica un cambio únicamente en el canal V, con lo cual frecuentemente no es utilizado por no acarrear información de color [59].

En el marco del problema del RALS la segmentación implica determinar la posición de las manos y el rostro en la escena. Para el caso de gestos dinámicos, al momento de determinar la posición de los rasgos de importancia en un *frame* dado, resulta conveniente considerar las detecciones de los *frames* anteriores dando

[10]Traducido del inglés, *saliency map*.

lugar al proceso de *seguimiento* –o *tracking*–. El proceso de *tracking* hace uso de la relación temporal existente entre las detecciones a los fines de mejorar el desempeño del sistema de reconocimiento [67].

2.4. Extracción de características

La extracción de características es un proceso fundamental para el reconocimiento automático. Mediante este proceso se realizan mediciones sobre las imágenes que brinden descripciones numéricas ligadas a los aspectos de interés para efectuar la tarea de clasifación. Un conjunto de características razonable debe satisfacer los siguientes requerimientos: (1) baja variabilidad intraclase: imágenes de objetos de la misma clase deben tener características numéricamente similares, (2) gran separación interclase: imágenes de objetos de clases diferentes deben poseer características notablemente diferentes, y (3) invarianza a cambios de escala, traslación y rotación, esto es, los objetos de interés deben ser descriptos de forma independiente de su tamaño, ubicación y orientación en la escena [71].

En el marco del RALS esta etapa consiste en la extracción de indicadores numéricos que brinden descripciones de la geometría de la mano segmentada, de los principales rasgos faciales de un sujeto o de la dinámica temporal de una seña. Si bien el problema del RALS ha sido frecuentemente tratado como un caso particular del reconocimiento de rasgos manuales, en la Sec. 1.1.1 se vio que en algunos casos los rasgos no manuales son tan importantes que sólo a partir de su consideración es posible determinar la semántica de la seña. A continuación, en la Sec. 2.4.1 se presentan distintos descriptores empleados para clasificar los rasgos manuales y en la Sec. 2.4.2 se introducen distintos abordajes para la consideración de los rasgos no manuales.

2.4.1. Rasgos manuales

En la Figura 2.15 se pueden observar algunas de las características computadas a los fines de describir la geometría del *blob* de cada mano, obtenido mediante el proceso de segmentación.

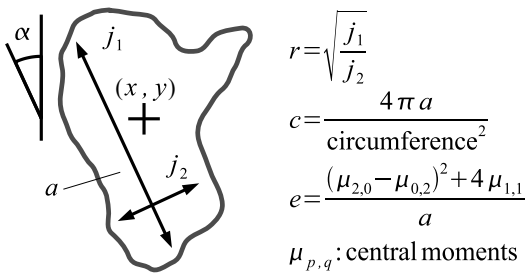


Figura 2.15: Algunas características de la geometría de cada mano. Tomada de [134].

Las características mostradas en la Figura 2.15 se describen brevemente a continuación [67, 121]:

- (x, y) son las coordenadas del centroide del *blob*.
- a es el área del *blob*, la cual puede medirse en cantidad de píxeles.
- α es la orientación del eje principal respecto a una dirección de referencia, en este caso, la vertical.
- r es la relación entre los momentos de inercia, siendo j_1 y j_2 los momentos a lo largo y perpendicular al eje principal, respectivamente [67]. Esta característica mide qué tan esbelto es el *blob*.
- c se denomina *compacidad* –o *compactness*– y es una medida de qué tan cerca del centro del *blob* se encuentran los píxeles. Teóricamente, $0 < c \leq 1$, siendo igual a 1 para el caso de un *blob* circular [20].
- e se denomina *excentricidad* –o *eccentricity*– y es una medida de qué tan lejos está el *blob* de tener simetría rotacional alrededor de su centroide [20].

En el caso de gestos dinámicos, la derivada primera de la posición (x, y) del centroide brinda una medida de la dinámica temporal del movimiento de la mano, así como la derivada primera del área lo hace con respecto al cambio de forma [67]. A modo de ejemplo, el área a es una característica útil pero fuertemente dependiente de la resolución de la imagen y de la distancia entre el señante y la cámara [67]. Además, las coordenadas (x, y) del centroide dependen de la posición del señante en la escena. Luego, resulta conveniente llevar a cabo la normalización de estas características de modo que el reconocimiento sea más robusto frente a cambios de escala y posición. Por esta razón suele ser frecuente, referir las posiciones detectadas a un punto anatómico específico, como podría ser el rostro o el hombro correspondiente a la mano segmentada [35]. Asimismo, una solución parcial al problema de cambio de escala es normalizar el área de la mano con respecto al área del rostro [67]. Éstos son aspectos cruciales en el desarrollo de sistemas independientes del señante. En este sentido, el uso independiente de sensores específicos de la actividad de las manos, tales como Leap Motion o guantes de datos, se ve seriamente limitado, razón por la cual es frecuente su uso en forma conjunta con un sistema que capture la actividad del cuerpo del señante [87].

Además de las características de la Figura 2.15 en la literatura se ha reportado el uso de características basadas en cerco convexo [59], SIFT [109] y transformada de Radon [111].

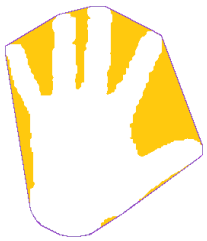
El *cerco convexo* o la *envolvente convexa*[11] H de un *blob* S se define como el conjunto convexo más pequeño que contiene a S [59]. En la Figura 2.16a se ilustra una mano segmentada en el espacio de color HSV [123]. En la Figura 2.16b en color *violeta* se muestra el cerco convexo de la mano. Además, en color *amarillo* se muestra el *defecto de convexidad*[12], definido como la diferencia $H - S$ [59].

[11] Traducidos del inglés, *convex hull*.

[12] Traducido del inglés, *convex deficiency* [59].

Estas características y medidas derivadas, por ejemplo la distancia euclidea entre S y H conforme se recorre la frontera de H, han mostrado ser muy útiles para el reconocimiento de configuraciones manuales y la identificación de la punta de los dedos, incluso de forma dinámica [29, 81, 127, 129]. A modo ilustrativo, en la Figura 2.16b puede verse que la distancia euclidea entre S y la frontera de H mostrará un máximo local en cada uno de los vértices interdigitales de la mano.

(a) Mano segmentada.



(b) Cerco convexo (violeta) y defecto de convexidad (amarillo).

Figura 2.16: Cerco convexo y defecto de convexidad de una mano segmentada [123]. Fuente de la imagen cruda: base de datos LSA16 [111].

Por su parte, la transformada SIFT[13] extrae una serie de descriptores de la geometría local en ciertos puntos de interés, para lo cual se efectúan dos operaciones sucesivas e independientes [109]: (1) detección de *puntos relevantes* –en adelante, *keypoints*– mediante un abordaje multiescala en el espacio *diferencia de gaussianas* y (2) extracción de un descriptor de la geometría local de cada *keypoint*. Los descriptores SIFT han probado ser robustos frente a una amplia familia de transformaciones, tales como cambios de escala y rotación, pequeños cambios de perspectiva, ruido, cambios de iluminación y contraste y deformación geométrica de la escena, con lo cual han sido frecuentemente empleados en el problema del reconocimiento automático de objetos –u *object recognition*– [109]. Este descriptor ha sido empleado para el reconocimiento de gestos manuales, tanto estáticos como dinámicos [38, 111, 128]. En el método de SIFT la detección de los *keypoints* es realizada de forma automática a partir de una representación multiescala de la imagen [109]. Luego, la cantidad de *keypoints* obtenidos para cada imagen puede ser *a priori* diferente. Surge así la alternativa de imponer la posición de los *keypoints*, dando lugar a una variante denominada DenseSIFT [51, 130]. Existen a su vez variantes de SIFT que buscan optimizar el costo computacional. Una de ellas se denomina SURF[14], fue introducida en 2006 [17] y empleada para el reconocimiento de gestos manuales dinámicos en [63]. Otra de ellas, 2 órdenes de magnitud más rápida que SIFT, se denomina ORB[15] fue introducida en [113] y se empleó en [80] para la detección de manos en un sistema basado en video en primera persona, bajo variaciones de posturas manuales, de iluminación y de fondo.

[13] Por sus siglas en inglés, *Scale Invariant Feature Transform*.
[14] Por sus siglas en inglés, *Speeded-Up Robust Features*.
[15] Por sus siglas en inglés, *Oriented fast and Rotated Brief*.

La transformada de Radon de una imagen 2D puede definirse como la integral de línea sobre la imagen a lo largo de una recta dada por (b, θ), donde b es la distancia perpendicular de la recta al origen y θ es el ángulo con respecto al eje horizontal de la imagen. Luego, es posible representar la imagen en el dominio (b, θ) [111]. En la Figura 2.17 se observa la transformada de Radon de una mano segmentada en el plano (b, θ) [123]. Resulta muy útil para la descripción de configuraciones manuales, sobre todo si se emplea el contorno de la mano [110].

(a) Imagen de entrada. (b) Entrada en escala de grises. (c) Transformada de Radon de la entrada.

Figura 2.17: Transformada de Radon de la mano segmentada [123]. Fuente de la imagen cruda: base de datos LSA16 [111].

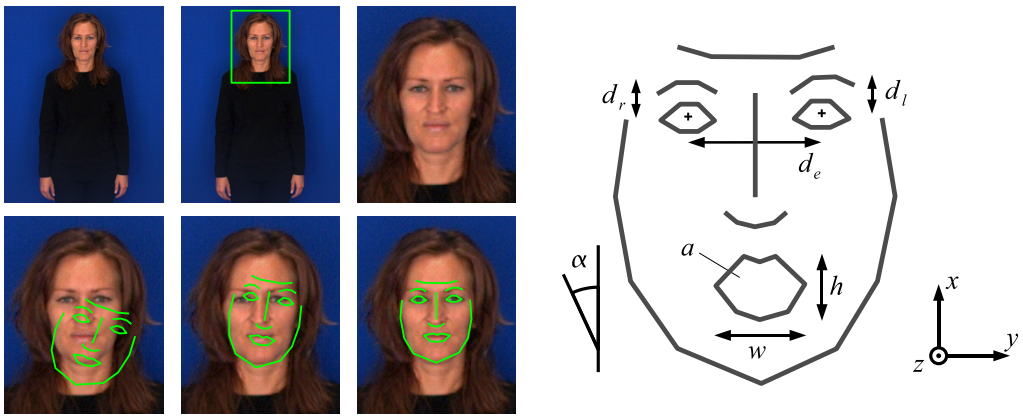
2.4.2. Rasgos no manuales

Desde sus comienzos los sistemas de RALS se han basado exclusivamente en los rasgos manuales [134]. No obstante, tal como fue mencionado en la Sec. 1.1.1, los rasgos no manuales son de suma importancia en la lengua de señas, tanto para comprender la gramática de la lengua como para determinar el significado de una seña aislada. De esta manera, resulta natural pensar que un sistema robusto de RALS también debería considerar estos aspectos. En este sentido, existen diversos trabajos que toman en consideración los rasgos no manuales para el RALS, tanto la actividad del rostro como la postura del cuerpo y la cabeza [13, 67, 88, 134].

De todos los rasgos no manuales, quizás el más importante sea la expresión facial, que incluye la mirada, la postura de las cejas y los patrones labiales. En esta línea, Von Agris y cols. fueron los primeros en estudiar en profundidad el impacto de los rasgos faciales sobre el desempeño de un sistema de RALS en distintos escenarios, logrando mejoras de hasta el 7 % en el reconocimiento de discurso continuo con independencia del señante y mostrando que en general la tasa de reconocimiento mejora al adicionar los rasgos faciales [134].

El análisis y la interpretación de la expresión facial requiere la extracción de los objetos de interés, tales como los ojos, las cejas y la boca, así como también las relaciones espaciales existentes entre éstos [134]. Para ello, es preciso segmentar el rostro del resto de la escena y sobre el mismo localizar los objetos de interés. Estos dos procesos pueden realizarse de forma simultánea en el caso que la detección del rostro esté basada en las características geométricas buscadas [149]. Tal es el caso de la caracterización del rostro mediante un *modelo de apariencia activa*

(AAM[16]). El AAM fue introducido en [48] y consiste de dos componentes: un modelo estadístico descriptivo de la apariencia de un objeto buscado –en este caso, un rostro humano– y un algoritmo para el *matching* de este modelo a un ejemplo de tal objeto en una escena dada. En la Figura 2.18 se ilustra el método seguido por Von Agris y cols. [134]. En primer lugar, se identifica la región del rostro y se realiza un *cropping* y un reescalado de ésta (parte superior de Figura 2.18a). En segundo lugar, se realiza el *matching* de un grafo predefinido de rostro (parte inferior de Figura 2.18b). Una vez que el grafo ha sido "ajustado" al rostro del señante, es posible calcular las características que se muestran en la Figura 2.18b y considerarlas luego para la clasificación [134].



(a) *Cropping* del rostro y *matching* del AAM. (b) Características del grafo "ajustado".

Figura 2.18: Interpretación de la expresión facial mediante AAM. Tomadas de [134].

Yacoob y Davis [142] emplearon la dinámica temporal de ciertas regiones faciales para reconocer expresiones faciales o gestos. En particular, ellos calcularon el flujo óptico sobre descriptores del movimiento local, para determinar cuáles son las zonas del rostro que participan en la generación de cada una de las 6 expresiones consideradas –felicidad, tristeza, sorpresa, enojo, miedo y disgusto–.

Es preciso aclarar que no todas las expresiones faciales tienen sentido en el contexto de una lengua de señas. A diferencia de los gestos faciales espontáneos desencadenados por una cierta emoción, los rasgos faciales de una lengua de señas son de uso deliberado. Ming y Ranganath [88] exploraron el reconocimiento automático de una serie de expresiones faciales –3 de la parte superior del rostro y 3 de la parte inferior– asociadas comúnmente a las lenguas de señas. En particular, trabajaron sobre la identificación de 3 tipos de preguntas analizando el comportamiento de las cejas y los hombros, y el uso de 3 modificadores sintácticos analizando la dinámica de los labios. Partiendo de imágenes de rostros previamente segmentados, Ming y Ranganath emplearon *Independent Component Analysis* y *Gabor Wavelet Networks* para la representación de las 6 expresiones faciales mecionadas [88]. En particular, hallaron tasas de reconocimiento superiores al 85 %

[16]Por sus siglas en inglés, *Active Appearance Model*.

empleando *Gabor Wavelet Networks*, tanto para la identificación de los tres tipos de preguntas como de los 3 modificadores sintácticos considerados.

Dentro de los rasgos no manuales de las señas se ha considerado además la descripción de la postura corporal como una característica de interés para el reconocimiento. Tal es el caso de los trabajos de Konstantinidis y cols. [76] y Mocialov y cols. [90], los cuales se comentarán con mayor detalle en la Sec. 2.6.

2.5. Clasificación

La etapa de clasificación de un sistema de RALS tiene como objetivo traducir las características de entrada a información categórica, y de este modo interpretar la semántica de los datos sensados. En otras palabras, la etapa de clasificación de un sistema de RALS permite reconocer el mensaje presente en los datos de entrada. Dependiendo del tipo de gesto que se desee clasificar, la etapa de clasificación empleará estrategias distintas. Para el caso de gestos estáticos, será suficiente contar con un clasificador capaz de asociar correctamente un vector de características de entrada a una clase. No obstante, para clasificar un gesto dinámico, no sólo será necesario considerar la actividad en cada *frame* entrante de manera independiente, sino también su evolución a lo largo del tiempo. Un abordaje muy empleado para la clasificación de gestos dinámicos, tanto para señas aisladas como para discurso continuo, es el reconocimiento basado en sub-unidades.

Los métodos de aprendizaje automático pueden ser agrupados en dos grandes familias: métodos de aprendizaje supervisado y métodos de aprendizaje no supervisado. Los métodos de aprendizaje supervisado se basan en una etapa de entrenamiento en la cual el sistema aprende a clasificar ciertos patrones en los datos de entrada, a partir de un conjunto de datos de entrada y sus correspondientes salidas deseadas [21]. Normalmente se refiere a este conjunto de datos como conjunto de datos de entrenamiento. Por otra parte, los métodos de aprendizaje no supervisado son usados para describir la estructura de los datos de entrada en función de sus atributos. Un ejemplo de este tipo de métodos lo constituyen los algoritmos de *agrupamiento* –o *clustering*– [21].

Tanto para la clasificación de gestos estáticos como dinámicos, el desempeño de un método de RALS se evalúa según su *tasa de reconocimiento* –también conocida como *accuracy*–, esto es, la proporción de clasificaciones correctas que realiza el sistema, independientemente de la clase. No obstante, para el caso de clasificar secuencias de señas, existen medidas de desempeño más sofisticadas, tales como la *word error rate* que toman en cuenta la cantidad de omisiones, la cantidad de inserciones y la cantidad de sustituciones realizadas [47].

2.5.1. Clasificación de gestos estáticos

El reconocimiento de gestos estáticos es realizado a partir de imágenes aisladas. Frecuentemente se ha empleado para el reconocimiento del alfabeto dactilológico de una lengua de señas dada [31].

Análisis discriminante lineal. La idea esencial detrás de este método de clasificación –en adelante, LDA[17]– es encontrar una combinación lineal de las características de entrada de modo de obtener la máxima discriminación entre las clases implicadas. Esto puede lograrse calculando una matriz de proyección óptima que, simultáneamente, maximice la separación entre las medias de las clases proyectadas y, al mismo tiempo, minimice la varianza de cada clase, reduciendo de esta manera el solapamiento entre las clases en el nuevo espacio [21]. Luego, frente a un dato entrante la clasificación se realiza preguntando en qué lado de la frontera se encuentra el dato proyectado. Con este objetivo y considerando un problema multiclase, la función de costo J a maximizar es [21]:

$$J(\mathbf{w}) = \mathrm{Tr}\left\{\left(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^T\right)^{-1}\left(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^T\right)\right\}, \tag{2.1}$$

donde $\mathbf{W}$ es la matriz de proyección deseada, $\mathbf{S}_{\mathrm{W}}$ y $\mathbf{S}_{\mathrm{B}}$ son las matrices de covarianza intraclase e interclase, respectivamente, y $\mathrm{Tr}\{\cdot\}$ denota la traza de la matriz correspondiente. Para una descripción más detallada de la derivación de la expresión (2.1), por favor diríjase a [21].

Máquinas de soporte vectorial. La idea esencial detrás de estos clasificadores –en adelante, SVM[18]– es encontrar los parámetros de un hiperplano que maximice el margen de separación entre las clases implicadas. Luego, la clasificación se realiza preguntando "de qué lado" del hiperplano se encuentra el dato de entrada [21].

Formalmente, dado un vector de características de entrada $\mathbf{x}_i$, es posible realizar su clasificación a partir del $\mathrm{sgn}\{y(\mathbf{x}_i)\}$, siendo:

$$y(\mathbf{x}_i) = \mathbf{w}^T\phi(\mathbf{x}_i) + b, \tag{2.2}$$

donde $\phi(\cdot)$ es un *kernel* de transformación de las características a un espacio de dimensión superior y $\{\mathbf{w}, b\}$ los parámetros que determinan el hiperplano de separación.

A continuación, se formulará el problema de determinar los parámetros $\{\mathbf{w}, b\}$ para el caso de dos clases. Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ el conjunto de N vectores de entrada con sus correspondientes etiquetas $t_1, t_2, \ldots, t_N$, donde $t_n \in \{-1, 1\}$. Suponiendo que los datos son linealmente separables, es posible encontrar un hiperplano óptimo $\{\mathbf{w}, b\}$ tal que $t_n y(\mathbf{x}_n) > 0$, para $n = 1, 2, \ldots, N$. A partir de un análisis geométrico, es posible demostrar que la distancia perpendicular de un punto $\phi(\mathbf{x}_n)$ al hiperplano viene dada por [21]:

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n\left(\mathbf{w}^T\phi(\mathbf{x}) + b\right)}{\|\mathbf{w}\|}. \tag{2.3}$$

El *margen* de un clasificador se define como la distancia perpendicular entre la frontera de decisión y el dato más cercano, cualquiera sea su clase. Dado que

[17] Por sus siglas en inglés, *Linear Discriminant Analysis*.
[18] Por sus siglas en inglés, *Support Vector Machines*.

en una SVM se busca maximizar el margen, el problema de optimización puede escribirse como [21]:

$$\underset{\mathbf{w},\,b}{\text{argmáx}} \left\{ \frac{1}{\|\mathbf{w}\|} \underset{n}{\text{mín}} \left[t_n \left(\mathbf{w}^T \phi(\mathbf{x}_n) + b \right) \right] \right\}. \tag{2.4}$$

No obstante, tal como se ha planteado el problema resulta muy difícil de resolver, para lo cual es posible hacer uso del hecho de que si $\mathbf{w}$ y b son las soluciones que brindan el margen máximo, entonces $\kappa\mathbf{w}$ y κb –siendo κ una constante– también lo serán. Luego, sin pérdida de generalidad, es posible imponer que la solución óptima satisfaga la restricción [21]:

$$t_n \left(\mathbf{w}^T \phi(\mathbf{x}) + b \right) \geq 1, \text{ para } n = 1, 2, \ldots, N. \tag{2.5}$$

En particular, la igualdad de la expresión anterior se cumplirá para el punto más cercano a la frontera de decisión. Es posible notar que el mínimo de la expresión (2.4) será 1 para algún valor de n. Luego, el problema consiste en hallar $\{\mathbf{w}, b\}$ que maximice $\|\mathbf{w}\|^{-1}$, sujeto a la restricción dada en la expresión (2.5). De forma equivalente, el problema puede reformularse como [21]:

$$\underset{\mathbf{w},\,b}{\text{argmín}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\}, \text{ sujeto a } t_n(\mathbf{w}^T \phi(\mathbf{x}) + b) \geq 1, \text{ para } n = 1, 2, \ldots, N. \tag{2.6}$$

Empleando multiplicadores de Lagrange, la función de costo J resulta [21]:

$$J(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n \left(\mathbf{w}^T \phi(\mathbf{x}) + b \right) - 1 \right\}, \tag{2.7}$$

donde $\mathbf{a} = (a_1, a_2, \ldots, a_N)^T$ y $a_n \geq 0$ para $n = 1, 2, \ldots, N$. El signo negativo del segundo término implica que se minimizará con respecto a $\{\mathbf{w}, b\}$ al mismo tiempo que se maximizará con respecto a $\mathbf{a}$.

El problema de una SVM puede generalizarse para datos que no son linealmente separables, para lo cual se introduce un conjunto de N constantes $\xi_n \geq 0$, denominadas *variables de holgura* –o *slack variables*– que permiten "relajar" las restricciones sobre el margen para cada dato. Dado un dato $\mathbf{x}_n$, la función ξ_n se define como [21]:

$$\xi_n = \begin{cases} 0, & \text{para } \mathbf{x}_n \text{ tal que } t_n y(\mathbf{x}_n) > 0 \\ |t_n - y(x_n)|, & \text{para } \mathbf{x}_n \text{ tal que } t_n y(\mathbf{x}_n) \leq 0 \end{cases} \tag{2.8}$$

Luego, para $n = 1, 2, \ldots, N$, el problema de optimización puede escribirse como [21]:

$$\underset{\mathbf{w},\,b}{\text{argmín}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^{N} \xi_n \right\}, \text{ sujeto a } t_n(\mathbf{w}^T \phi(\mathbf{x}) + b) \geq 1 - \xi_n. \tag{2.9}$$

En este caso la función de costo resulta [21]:

$$J(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\left\{t_n\left(\mathbf{w}^T\phi(\mathbf{x}) + b\right) - (1-\xi_n)\right\} - \sum_{n=1}^{N}\mu_n\xi_n, \tag{2.10}$$

donde $\{a_n \geq 0\}$, $\{\mu_n \geq 0\}$ son los multiplicadores de Lagrange y $C > 0$ es un parámetro de ponderación entre el margen y la penalización de las variables de holgura. Para mayor detalle, por favor diríjase a [21].

La clasificación de un dato de entrada en más de 2 clases mediante una SVM puede abordarse como múltiples problemas binarios de clasificación, tanto bajo el esquema "uno *versus* uno" como "uno *versus* el resto", o bien a partir de la construcción de una SVM multiclase real [65]. Para una descripción más detallada de la formulación matemática de estas variantes, por favor diríjase a [65].

En [128] Tharwat y cols. emplearon este clasificador para el reconocimiento de 30 configuraciones manuales de la Lengua de Señas Árabe. Como vector de características emplearon los descriptores SIFT, los cuales fueron reducidos de dimensión con LDA. La base de datos empleada se compuso de 210 imágenes de 200×200 en escala de grises, 7 muestras de 30 caracteres de la Lengua de Señas Árabe, obtenidas bajo diferentes condiciones de iluminación, orientación y grado de oclusión. Este sistema mostró tasas de reconocimiento del orden del 98 %. Los experimentos realizados muestran una gran robustez de las características frente a cambios de escala y orientación de las imágenes de entrada, así como también frente a cambios en el número de ejemplos usados para el entrenamiento. Incluso los autores también estudiaron la robustez del método frente a oclusiones en las imágenes de entrada, mostrando caídas en el desempeño menores al 5 % para el 40 % de oclusión.

En [103] Pisharady y cols. proponen un método para la segmentación y el reconocomiento de posturas manuales bajo fondos complejos y naturales. Para ello, hacen uso de un modelo bayesiano de atención visual a partir del cual realizan la segmentación. La clasificación de las posturas manuales se efectúa en base a características de forma y textura de la región de la mano mediante un clasificador SVM, con el cual obtienen una tasa de acierto superior al 94 % en la clasificación de 10 posturas manuales distintas en diferentes contextos de fondo.

En [145] se presenta un sistema para la clasificación de configuraciones manuales a partir de modelos activos de forma basados en SVM. Asimismo, en [69] se presenta un sistema independiente del señante para el reconocimiento de posturas manuales.

K vecinos más cercanos. El aprendizaje de estos clasificadores consiste en asignar a cada dato del conjunto de entrenamiento la clase a la cual pertenece. Luego, dado un dato a clasificar, empleando alguna medida de distancia –por ejemplo, la distancia euclidea– en primer lugar deben encontrarse sus *K vecinos más cercanos*. Finalmente, el dato de entrada es clasificado con la etiqueta con más repeticiones o votos dentro de los K vecinos considerados [21]. Debido a que se trata de un método no-paramétrico, frecuentemente funciona muy bien en situaciones donde

la frontera de decisión es muy irregular [21]. En adelante, este método será referido como 'clasificador/clasificación por KNN[19]'.

En [61] Gupta y cols. realizaron reconocimiento de 26 caracteres del alfabeto de la Lengua de Señas India, el cual se compone tanto de posturas unimanuales como bimanuales. Para ello, Gupta y cols. propusieron en primera instancia una clasificación en señas unimanuales o bimanuales. Luego, para la identificación de cada carácter emplearon una combinación de SIFT y HOG[20] como características. Como etapa de clasificación, emplearon un clasificador por KNN, logrando tasas de reconocimiento de 97.5 % y de 91.1 % para las posturas unimanuales y bimanuales, respectivamente.

Se han realizado varios trabajos donde se compara la tasa de acierto de SVM *versus* KNN, bajo mismo tamaño de datos entrenamiento y testeo. En general, se ha observado que el desempeño de KNN es más pobre [32, 79, 128].

Redes neuronales. Las *redes neuronales* –en adelante, NN[21]– son sistemas inspirados en la estructura y fisiología del tejido nervioso, muy empleados en la inteligencia artificial desde mediados del siglo pasado. A partir del aprendizaje supervisado, la idea principal de las redes neuronales es aproximar una función que relacione patrones presentes en los datos de entrada con una etiqueta de salida deseada [110]. Una NN se compone por capas de neuronas: una capa de entrada, una o más capas ocultas y una capa de salida. A su vez, cada capa se encuentra compuesta por una cantidad arbitraria de unidades funcionales, cada una de ellas denominada *neurona* o *unidad*. La salida y de una neurona viene dada por [21]:

$$y = h\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{2.11}$$

donde x_i es la $i-$ésima entrada, w_i es el $i-$ésimo peso, b es un parámetro de *bias* y $h(\cdot)$ es una función de activación no lineal. Luego, el comportamiento de una red se define en base a tres aspectos: los patrones de interconexión entre las distintas capas de neuronas, los pesos de las interconexiones y la función de activación empleada [21]. La arquitectura de una NN abarca desde el perceptrón simple hasta las redes profundas propias de los sistemas de *aprendizaje profundo* –o *deep learning*–, las cuales serán comentadas con mayor detalle en la Sec. 2.6.

En [10] Al-Jarrah et al. presentaron dos NNs para el reconocimiento de 30 caracteres del alfabeto manual de la Lengua de Señas Árabe. Para ello partieron de imágenes estáticas de manos sin guante, bajo fondo uniforme y 60 muestras por clase. El sistema fue alimentado mediante descriptores de la configuración de la mano en la imagen, en particular las distancias existentes entre el centroide y el borde de la mano en ciertas orientaciones preestablecidas. De forma previa a la extracción de características, los autores realizan una normalización de la

[19] Por sus siglas en inglés, *K-Nearest Neighbours*.

[20] Por sus siglas en inglés, *Histogram of Oriented Gradients*. Mediante HOG se cuantifica la ocurrencia relativa de ciertas orientaciones del vector gradiente, a partir de la cual es posible describir la geometría local de una imagen.

[21] Por sus siglas en inglés, *Neural Networks*.

orientación de la mano, de modo que la firma obtenida sea invariante a rotaciones, además de ser invariante a traslaciones por estar referidas al centro de la mano. Para la clasificación usaron un tipo de red particular llamada ANFIS, alcanzando una tasa de acierto del 93.5 %. En trabajos posteriores en la misma línea, los autores reportaron una tasa de acierto del 97.5 % [9].

En [16] proponen un sistema para el reconocimiento de 40 gestos manuales estáticos componentes de la Lengua de Señas Brasilera, comúnmente referida como LIBRAS[22]. La base de datos empleada contiene 240 repeticiones de cada gesto, lo cual redunda en 9600 imágenes RGB. Bastos y cols. emplearon HOG y ZI como descriptores los bordes y las formas presentes en las imágenes de las manos segmentadas. Agruparon los datos en 12 grupos y realizaron el reconocimiento mediante NNs en un esquema de dos etapas: (1) reconocimiento del grupo al cual pertenece la imagen de entrada; (2) reconocimiento del gesto manual dentro del grupo. En particular, los autores emplearon un perceptrón multicapa por cada grupo. La salida de cada uno de estos a su vez alimenta una NN particular para la identificación del gesto propiamente dicho. El abordaje en dos etapas sucesivas busca reducir el costo computacional y el crecimiento excesivo de la red. No obstante, una clasificación errónea de grupo impedirá la posibilidad de clasificar de forma correcta. Finalmente, los autores reportaron una tasa de reconocimiento promedio del 96.7 % y de un 86 % en el peor de los casos.

2.5.2. Clasificación de gestos dinámicos

En caso de querer clasificar gestos dinámicos, además de encontrar patrones espaciales en cada imagen, será necesario detectar y eventualmente asociar patrones en las series temporales dadas por las características de estas imágenes. Dos técnicas muy empleadas en este sentido son los Modelos Ocultos de Markov y la Alineación Dinámica de Tiempo.

Modelos ocultos de Markov. La dinámica de un sistema puede caracterizarse mediante las respuestas temporales del mismo frente a entradas específicas, o bien a partir de un modelo del sistema en su espacio de estados [95]. Bajo este último enfoque, los *modelos ocultos de Markov* –en adelante, HMM[23]– constituyen una herramienta muy utilizada para modelar la dinámica de un sistema en su espacio de estados, sujeta a la particularidad de que los estados en consideración no son directamente observables; de allí su cualidad de *hidden* u oculto. Haciendo uso de un enfoque estadístico, un HMM se construye suponiendo que existe una distribución de probabilidad de observar un determinado símbolo en particular –variable observable– para cada estado posible del sistema. De manera formal, un HMM se determina a partir de [107]:

- El número de estados N entre los cuales puede evolucionar el sistema modelado.

[22] Por sus siglas en portugués, *Língua Brasileira de Sinais.*
[23] Por sus siglas en inglés, *Hidden Markov Models.*

- Un conjunto de coeficientes $\{a_{ij}\}$ que representan la probabilidad de que el sistema evolucione de un estado i a un estado j en el instante t. De forma simbólica, $a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$, con $1 \leq i, j \leq N$, donde s_i se emplea para hacer referencia al estado i y q_t para referirse al estado correspondiente al instante t. Normalmente estos coeficientes quedarán dispuestos en una matriz $A \in \mathbb{R}^{N \times N}$, a la cual denominaremos *matriz de transición de estados*.

- Un conjunto de símbolos v_k, con $k = 1, 2, \ldots, M$, asociados a cada uno de los estados.

- Un conjunto de coeficientes $\{b_j(k)\}$ que representan la probabilidad de observar el símbolo v_k dado que el sistema se encuentra en el estado j. De forma simbólica, $b_j(k) = P(v_k | q_t = s_j)$, con $1 \leq j \leq N$ y $1 \leq k \leq M$. Normalmente estos coeficientes se disponen en una matriz B de tamaño $M \times N$ que denominaremos *matriz de emisiones*.

- Una distribución de probabilidad de los estados iniciales $\pi = \{\pi_i\}$, donde $\pi_i = P(q_1 = s_i)$, con $1 \leq i \leq N$.

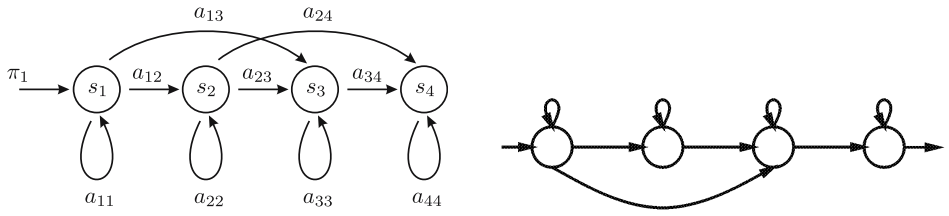
Un HMM permite modelar la dinámica de un sistema y responder a preguntas como "dado un modelo y una secuencia de observaciones, ¿cuál es la secuencia más probable de estados por los cuales evolucionó el sistema?", o "dado un modelo y una secuencia de observaciones, ¿cuál es la probabilidad de que esta secuencia de observaciones haya provenido del modelo dado?". Extendiendo esta idea resulta lógico pensar que si se tuvieran varios modelos y una secuencia de observaciones, comparando probabilidades sería posible responder la pregunta "¿cuál es el modelo que brinda la mayor probabilidad de haber generado tal secuencia de observaciones?" [107]. Bajo este último enfoque es posible llevar a cabo la clasificación mediante HMMs.

En cuanto al RALS, tomando en cuenta el buen desempeño de los HMMs para el *reconocimiento automático del habla* y su similitud con el problema del RALS, desde mediados de los años 90s los modelos ocultos de Markov se han utilizado ampliamente como etapa de clasificación de gestos dinámicos [34].

En el año 1995 [122] Starner expuso su trabajo de maestría en el cual reconoció oraciones en Lengua de Señas Americana a partir de la captura de gestos manuales mediante una sola cámara RGB, siendo uno de los trabajos pioneros en los sistemas de RALS basados en visión y en extender el reconocimiento de señas aisladas a oraciones. Starner adquirió un léxico de 40 palabras, conformado por señas de pronombres, verbos, sustantivos y adjetivos. Luego, construyó 494 oraciones concatenando señas aisladas escogidas aleatoriamente que cumplan con una estructura gramatical prestablecida. En particular, la estructura impuesta fue "pronombre personal, verbo, sustantivo, adjetivo y el mismo pronombre personal". Los sujetos participantes de la base de datos vestían guantes de color –amarillo en mano derecha, anaranjado en izquierda–. Las manos fueron segmentadas por *crecimiento de región* –o *region growing*–, fijando la semilla en un píxel por color adecuado. Luego, las características empleadas fueron el centroide del *blob* de cada mano, la orientación y la excentricidad de las elipses que mejor ajustan a cada *blob*.

Mediante HMMs, Starner obtuvo una tasa de reconocimiento del 91 %. Tomando en consideración la estructura gramatical empleada para la construcción de las oraciones durante la clasificación, Starner observó una tasa del 97 %, observando que los errores cometidos en este último caso sólo son debidos a susticiones, dado que no puede haber errores de omisión o inserción de palabra. En el año 1996, bajo un esquema de trabajo similar Starner y Pentland [120] reportaron una tasa de reconocimiento del 99 % empleando guantes de colores y del 92 % con manos sin guantes. [121]. Siguiendo la misma línea, en 1998 Starner y cols. [121] presentaron un sistema basado en visión en primera persona, con una cámara montada RGB en la visera de una gorra. Mediante este sistema los autores reportaron una tasa de reconocimiento del 97 % con manos sin guantes y sin información gramatical *versus* el 92 % del sistema de visión en segunda persona.

Si bien existen muchos tipos de HMMs, sólo algunos de ellos serán útiles para modelar señales cuyas características varían en el tiempo de forma sucesiva [67]. En la Figura 2.19 se observa una representación gráfica de dos HMMs que permite observar su topología. En particular, en la Figura 2.19a se muestra el modelo de Bakis, el cual ha sido ampliamente utilizado en el campo del *reconocimiento automático del habla*, por su propiedad de compensar diferentes velocidades de articulación [67] y de ser un modelo de tipo *izquierda–derecha* [107]. De forma consistente con la notación presentada, los círculos denotan los estados s_i posibles del modelo. Las flechas denotan las transiciones posibles entre estados. Éstas quedan caracterizadas por los coeficientes a_{ij}, siendo a_{ij} la probabilidad de que el sistema evolucione desde el estado s_i al estado s_j. Por su parte, π_1 representa la distribución de probabilidad de estados iniciales. En la Figura 2.19b se muestra la topología determinada experimentalmente y empleada por Starner y cols. en [120, 121].

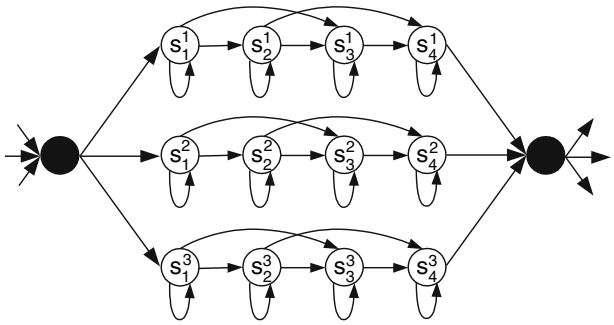


(a) Modelo de Bakis [67]. (b) HMM de 4 estados [120, 121].

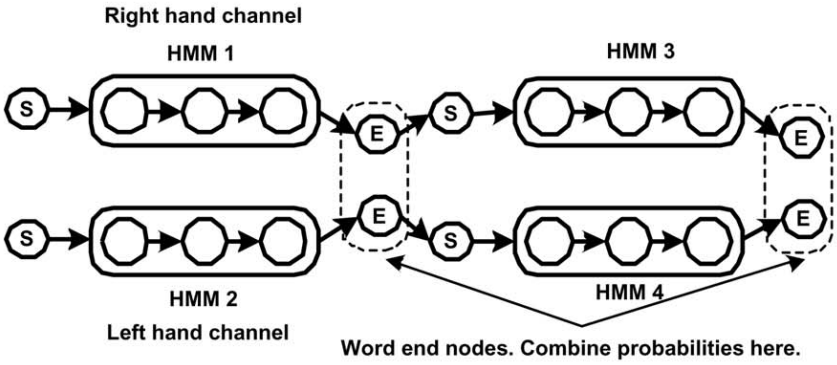
Figura 2.19: Diferentes topologías de HMMs.

Si bien los HMMs han mostrado ser útiles para modelar secuencias de datos, es necesario proponer variantes para la consideración de dinámicas *simultáneas* para la clasificación. Éste es el caso de los sistemas de RALS basados en más de un elemento sensor, o bien en aquellos sistemas basados en una cámara simple que consideran la dinámica de más de una región de la escena, por ejemplo, ambas manos, mano y rostro, manos y rostro. En este sentido se han propuesto diversas extensiones a los HMMs convencionales, entre las cuales se encuentran los HMMs paralelos [132] y los HMMs acoplados [22]. En la Figura 2.20 se muestran

dos ejemplos de HMMs paralelos (PaHMM[24]). En la Figura 2.20a se muestra el PaHMM de tres canales propuesto en [67]. En este modelo, cada una de las cadenas posee la topología de Bakis y la dinámica de cada canal es independiente del resto. En un estado particular denominado *estado confluyente* –círculo negro en la Figura 2.20a– se realizan las combinaciones de probabilidades de los diferentes canales resultando en una única probabilidad de la seña completa [67]. Por otro lado, en la Figura 2.20b se muestra el PaHMM empleado en [96] para la consideración simultánea de la dinámica de la mano izquierda y de la mano derecha.



(a) Tres canales, uno por rasgo [67].



(b) Dos canales, uno por mano [96].

Figura 2.20: Diferentes tipos de HMMs paralelos.

En [78] Kumar y cols. emplearon un HMM acoplado (CHMM[25]) para la fusión de los datos da salida de Leap Motion y Microsoft® Kinect™. A diferencia de los PaHMMs, en los CHMM los canales de información son interdependientes, esto es, los estados de los diferentes canales se influencian unos a otros. En la Figura 2.21 puede observarse la topología típica de un modelo CHMM. Obsérvese que si bien los estados se influencian unos a otros, las salidas continúan siendo independientes.

[24]Por sus siglas en inglés, *Parallel Hidden Markov Model*.
[25]Por sus siglas en inglés, *Coupled Hidden Markov Model*.

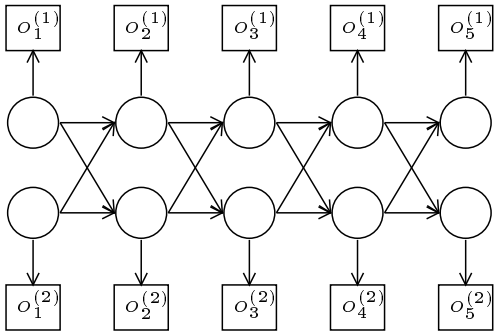


Figura 2.21: HMM acoplado. Tomada de [132].

Alineamiento temporal dinámico. El *alineamiento temporal dinámico* –en adelante, DTW[26]– es una técnica empleada para medir la similitud entre dos series temporales de diferente duración y dinámica, a partir de una expansión/compresión local en tiempo [19]. Sean $X = \{x_n\}_{n=1}^{N}$ e $Y = \{y_m\}_{m=1}^{M}$ dos series temporales de longitudes N y M, respectivamente. Las secuencias X e Y pueden relacionarse en un plano (n, m), donde cada punto (i, j) corresponde al alineamiento particular entre los puntos x_i e y_j, por medio de un mapeo no lineal W denominado *camino de alineación* –o *warping path*–. Para el cómputo de W, se construye una matriz de distancia de $N \times M$ entre todos los puntos de X y de Y y se busca el *camino* que minimiza la distancia global entre ambas series [19]. El mapeo W debe cumplir una serie de propiedades que garanticen cierto grado de preservación de los eventos en las series. Entre ellas, la *monotonicidad* –las secuencias $\{x_n\}$ y $\{y_m\}$ deben permanecer ordenadas en tiempo– y la *continuidad* –paso unitario en el espacio (n, m)–. Para más detalle, por favor diríjase a [19].

En relación a la clasificación de gestos dinámicos, esta técnica se utiliza para alinear por ejemplo las trayectorias de las manos entre una seña de prueba y una seña de referencia –o *template*–, viendo a ambas trayectorias como series temporales y clasificando la seña de prueba mediante un procedimiento de *matching* [110].

En [39] Darrell y Pentland proponen un método para el reconocimiento de gestos manuales dinámicos basado en visión. Haciendo uso de DTW, obtuvieron una tasa de reconocimiento del 96 % para el gesto correspondiente a 'hello' en Lengua de Señas Americana.

En [36] Corradini propone un método para el reconocimiento de un conjunto acotado de gestos dinámicos mediante DTW. En particular, Corradini estudió el reconocimiento de los gestos 'detenerse', mano izquierda o derecha saludando y mano izquierda o derecha señalando hacia la izquierda y derecha, respectivamente. Una vez localizado el sujeto en la escena, Corradini llevó a cabo la segmentación de las manos. Luego, extrajo un vector de 15 características de la configuración

[26]Por sus siglas en inglés, *Dynamic Time Warping*.

manual y las relaciones espaciales manos–cabeza, invariante a traslaciones y a cambios de escala. El vector de características se conformó a partir de los momentos centrales de orden 2 y 3 normalizados más la posición y velocidad normalizadas según tamaño de la imagen y con el origen de coordenadas en el centroide de la cabeza del sujeto. Luego, el autor seleccionó las secuencias de características representativas a comparar mediante DTW. Finalmente, mediante un clasificador por KNN sobre las medidas de distancia entre las secuencias post DTW, Corradini obtuvo tasas de reconocimiento entre el 88.5 % y el 91.6 % para los distintos gestos.

Lichtenauer y cols. [83] proponen la aplicación al RALS de una variante de DTW denominada DTW estadístico (SDTW[27]). La principal diferencia entre DTW y SDTW radica en que DTW realiza la alineación contra una señal determinística, mientras que SDTW realiza la alineación contra un modelo estadístico de la señal. En [83] los autores estudian el efecto de disociar el método de SDTW del proceso de clasificación. Asimismo, presentan dos clasificadores que denominan CDFDs y Q-DFFM y realizan una prueba de concepto sobre datos de movimientos manuales en 3D, poniendo en discusión aspectos teóricos sobre la etapa de clasificación. Entre otras, los autores concluyen que bajo este esquema es posible optimizar la clasificación, descartando las características *no discriminativas* post SDTW. Además, los esquemas secuenciales 'SDTW + CDFD' y 'SDTW + Q-DFFM' se desempeñan mejor que SDTW *per se*, e incluso mejor que HMM.

Celebi y cols. [27] proponen una mejora a una variante de DTW denominada 'DTW ponderado[28]'. En particular, Celebi y cols. trabajaron en el reconocimiento de gestos corporales empleando Microsoft® Kinect™ como elemento sensor. Como ya se explicó en la Sec. 2.2.1, en cada *frame* proveniente de Microsoft® Kinect™ se ajusta un modelo corporal de 20 nodos. La variante 'DTW ponderado' surge de la hipótesis de que no todos los nodos del modelo son igualmente importantes para la determinación de un gesto. En particular, la relevancia de cada nodo –y por lo tanto su ponderación– se define en términos de la contribución del mismo al patrón de movimiento global implicado en cada gesto particular. Los autores generaron una base de datos propia de 8 gestos corporales, con 28 muestras por gesto. Mediante el método propuesto reportaron una tasa de reconocimiento del 96 % *versus* el 62.5 % y el 60 % correspondientes a los métodos DTW ponderado original y DTW clásico, respectivamente.

En base a lo presentado en esta sección, la clasificación de una seña aislada puede llevarse a cabo a partir de un HMM convencional, o bien a partir de extensiones como PaHMM, la cual permite representar a la seña como una composición de rasgos simultáneos cuyas dinámicas son clasificadas individualmente y combinadas posteriormente para una clasificación de la seña completa [96]. Este último caso está fuertemente vinculado a la interpretación lingüística de una seña como una composición de sub-unidades. A continuación se introduce con mayor detalle este aspecto lingüístico y algunas propuestas de RALS bajo este enfoque.

[27]Por sus siglas en inglés, *Statistical Dynamic Time Warping*.
[28]Traducido del inglés, *weighted* DTW.

2.5.3. Reconocimiento basado en sub-unidades

Las lenguas de señas están compuestas por miles de señas. Los sistemas de reconocimiento de señas frecuentemente implementan un clasificador por gesto. No obstante, este abordaje se torna difícil de resolver prácticamente, incluso para tareas sencillas como reconocimiento de señas aisladas y su posterior búsqueda en diccionario [35]. Una estrategia para resolver este problema es reconocer cada seña a partir de las *sub-unidades* que la componen. En 1960, Stokoe estableció que una seña con significado puede interpretarse como una combinación de unidades subléxicas denominadas queremas [85, 124]. Los *queremas*[29] constituyen el correlato en lengua de señas de los fonemas en la voz hablada y pueden definirse como "el conjunto de posiciones, configuraciones y movimientos que funcionan de manera idéntica en el lenguaje", independientemente del contexto lingüístico [124]. El término 'querema' no ha sido ampliamente aceptado por la comunidad científica y resulta frecuente el empleo de los términos 'sub-unidad' o directamente 'fonema' para hacer referencia a este mismo concepto, ya sea para el análisis de la lengua o el abordaje de soluciones tecnológicas bajo este enfoque [35, 99, 108, 137].

El reconocimiento de una seña basado en sub-unidades se realiza en dos etapas. En primer lugar, se lleva a cabo una identificación de las sub-unidades de una seña o una oración. En segundo lugar, se lleva a cabo una combinación de las sub-unidades identificadas para la identificación de la seña. Bajo la hipótesis de que las señas comparten sub-unidades y que la cantidad de sub-unidades es limitada, es posible trabajar de forma más eficiente con léxicos de gran tamaño, siendo ésta la principal ventaja [35, 67, 137]. No obstante, las limitaciones en este sentido son la baja disponibilidad de datos con etiquetas a nivel de sub-unidad [67].

En [136] Waldron y Kim presentaron un método de una NN en dos etapas de clasificación para el reconocimento de señas aisladas de Lengua de Señas Americana. Las señas fueron sensadas mediante un guante de datos. Los datos fueron etiquetados y empleados para entrenar de forma supervisada las dos etapas de clasificación. En primer lugar, reconocen la seña a nivel fonológico, es decir las sub-unidades componentes. Luego, las sub-unidades reconocidas al comienzo, mitad y fin de la seña alimentan la segunda etapa de clasificación, la cual hace efectivo el reconocimiento de la seña propiamente dicha. En particular, Waldron y Kim reconocieron las siguientes sub-unidades: 36 configuraciones manuales, 10 ubicaciones, 11 orientaciones y 11 movimientos manuales. En particular, las configuraciones manuales, las ubicaciones y las orientaciones se reconocieron de forma instantánea en los momentos ya comentados. Los autores reportaron una tasa de reconocimiento del 84 % para un vocabulario de 14 señas provenientes de 6 sujetos.

Wang y cols. [137] proponen un método de reconocimiento basado en sub-unidades para el reconocimiento de Lengua de Señas China (CSL[30]). El abordaje empleado busca una solución que escale bien conforme aumenta el tamaño del léxico. El método propuesto reconoce 2400 sub-unidades de CSL mediante un

[29] Traducido del inglés, *cheremes*; proveniente del griego *cheir*, 'mano' [67]. La rama de la lingüística que estudia los queremas se denomina *querología*.

[30] Por sus siglas en inglés, *Chinese Sign Language*.

HMM para cada una. Luego, empleando una estructura de datos de tipo árbol y un modelo de lenguaje, lograron reconocer 5119 señas con una tasa del 92.8 %.

Aran y cols. [13] compararon varios métodos para fusionar las características manuales con las no manuales en un sistema de RALS. En particular, propusieron un clasificador en dos etapas secuenciales. La primera de ellas fue alimentada con los rasgos manuales y la segunda con los rasgos no manuales. La salida de la segunda etapa sólo fue considerada en casos de ambigüedad o duda, dando lugar a un esquema que denominaron *sequential belief-based fusion*. Bajo este esquema, Aran y cols. obtuvieron tasas de reconocimiento superiores al 81.6 %.

Almeida y cols. [11] propusieron un método para la extracción de características basada en la estructura fonológica de la Lengua de Señas Brasilera (LIBRAS). Para ello, emplearon un sensor de tipo RGB-D, extrajeron 7 características y estudiaron su relación con la configuración, el movimiento y la posición de las manos como elementos estructurales de toda lengua de señas. Haciendo uso de SVM para la clasificación, obtuvieron una tasa de reconocimiento promedio del 80 %, sobre un léxico de 34 señas de LIBRAS.

2.6. RALS via *aprendizaje profundo*

En [96] se reporta una serie de restricciones que comúnmente deben respetar los señantes al momento de hacer uso de un sistema de RALS basado en visión. Estas restricciones se traducen en hipótesis de los métodos empleados para el reconocimiento. Entre otras, se mencionan el uso de vestimenta manga larga por parte del señante, el uso de guantes de color por parte del señante, fondo uniforme o fondo complejo pero estacionario, cabeza/rostro con menor movimiento que las manos, movimiento constante de las manos, cuerpo fijo e introducción de la localización inicial de mano, mano no dominante y/o rostro excluidas del campo de visión, vocabularios restringidos o señas artificiales para evitar oclusiones, campo de visión restringido a las manos, las cuales se requieren a distancia y con una orientación fija con respecto a la cámara.

Con el objetivo de ir superando una a una estas restricciones, en los últimos años han comenzado a desarrollarse con mayor fuerza los sistemas de RALS basados en visión mediante técnicas de *aprendizaje profundo*[31]. En general, el aprendizaje profundo se basa en técnicas de aprendizaje supervisado. La particularidad de estos sistemas es que aprenden a clasificar directamente desde los datos, sin la necesidad de una etapa previa de extracción de características, siempre y cuando la cantidad de ejemplos de entrenamiento sea lo suficientemente grande. Dentro de los sistemas de aprendizaje profundo de uso masivo se encuentran las redes neuronales convolucionales y las redes neuronales recurrentes.

Por un lado, las *redes neuronales convolucionales* –en adelante CNN[32]– constituyen redes de tipo *prealimentadas* –o *feed forward*– normalmente empleadas para el reconocimiento de patrones en imágenes aisladas [60]. Por otro lado, las *redes*

[31]Rama de la inteligencia artificial popularmente conocida como *deep learning*.
[32]Por sus siglas en inglés, *Convolutional Neural Networks*.

neuronales recurrentes –comúnmente referidas como RNN[33]– poseen vías de realimentación de la representación interna de la información y es común que en su funcionamiento tomen no sólo la entrada actual sino también entradas pasadas. Luego, a diferencia de una CNN, una RNN posee la capacidad de memorizar internamente un estado resultante de procesos pasados. Por esta razón, las RNNs son comúnmente empleadas para el procesamiento y la clasificación de datos a nivel de secuencia [60]. En el marco del RALS, una RNN resultará de utilidad para la clasificación de una seña a partir del video completo, mientras que una CNN sólo permitirá clasificar los rasgos de la seña a nivel de *frame* aislado.

Gebre y cols. [57] propusieron un método para la discriminación de seis lenguas de señas distintas: británica, danesa, francesa-belga, flamenca, griega y holandesa. En primera instancia, a partir de pequeñas muestras de video de cada lengua escogidas aleatoriamente *aprendieron las características* de forma no supervisada. La base de datos empleada se compuso por videos provenientes de 30 señantes, balanceada en cuanto a la cantidad de muestras por lengua. Luego, emplearon estas características bajo un esquema supervisado para la identificación automática de las lenguas de señas consideradas, alcanzando una *accuracy* promedio de 84 %.

Koller y cols. [74] propusieron un sistema para el reconocimiento robusto de configuraciones de la boca típicas en las lenguas de señas mediante CNNs. Para el entrenamiento de la CNN hacen uso de una estrategia denominada *supervisión débil* –o *weak supervision*–, la cual consiste en relajar la condición de que cada *frame* de las muestras de video del conjunto de entrenamiento cuente con una *etiqueta* –o *label*– explícita asociada. Las salidas de la CNN son introducidas a un HMM que realiza un alineamiento temporal forzado y permite iterar el aprendizaje de la CNN a partir del video [74]. Los autores reportan una mejora en la clasificación de configuración de la boca del 8 % en el contexto del RALS.

Empleando un esquema de trabajo similar que en [74], en 2016 Koller y cols. [75] presentaron un sistema basado en una CNN denominado 'Deep Hand', para la clasificación de 60 configuraciones manuales propias de la DGS. Este mismo grupo de trabajo también ha presentado sistemas tanto para el reconocimiento de discurso continuo como para la traducción de lengua de señas a lengua escrita [33, 75]. Este último problema es aún más complejo que el reconocimiento de discurso continuo, puesto que deben tomarse en consideración las estructuras lingüísticas y gramaticales que brindan frases correctas desde el punto de vista semántico. Como se verá en el Capítulo 4, durante este trabajo se realizó un procesamiento y clasificación a nivel de *frame* aislado, para lo cual se empleó la red Deep Hand, la cual será descripta posteriormente con mayor detalle.

En base a los trabajos [25, 119, 139] se creó la librería gratuita 'OpenPose', accesible en https://github.com/CMU-Perceptual-Computing-Lab/openpose, que no sólo extrae la postura de las manos, sino también del cuerpo y del rostro, sin la necesidad de marcadores sobre el señante y permitiendo incluso la presencia de más de un individuo en la escena bajo condiciones de fondo variable. En la Figura 2.22 se muestra la salida de OpenPose, haciendo uso tanto del módulo para detección de postura manual como de la actividad facial.

[33] Por sus siglas en inglés, *Recurrent Neural Networks*.

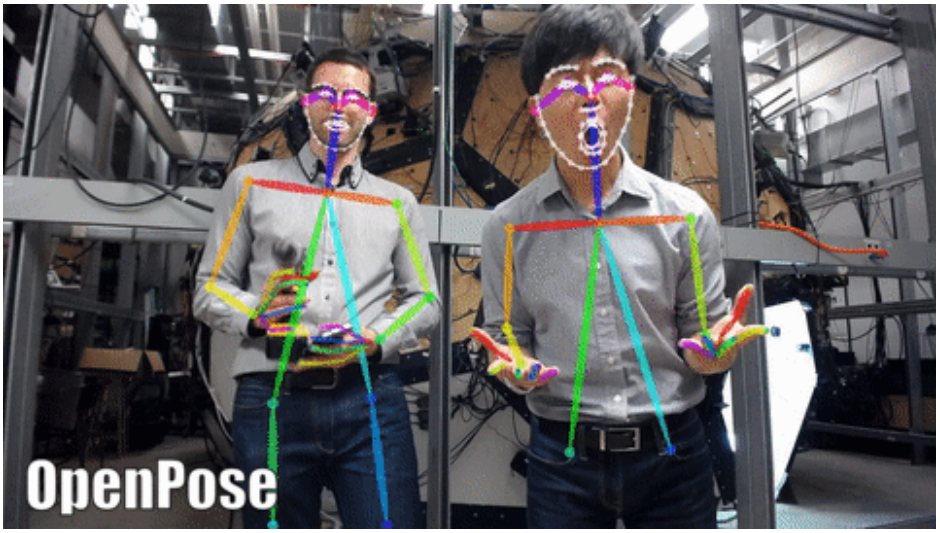


Figura 2.22: Muestra de la salida completa de OpenPose: detección de las posturas corporal, manual y facial. Tomada de [1].

El uso de OpenPose aplicado al RALS ya ha sido reportado en [76,90]. En [76] proponen el uso de un *sistema de meta–aprendizaje* –o *meta-learner*–, a partir del cual alcanzan una tasa de reconocimiento de seña aislada del 98 %, sobre una base de datos balanceada compuesta por 64 señas y 3200 videos provenientes de 10 sujetos. En [90] alcanzan tasas de reconocimiento del 80 % sobre video continuo y superiores al 95 % sobre señas previamente segmentadas en tiempo.

Por último, en 2018 Khan y cols. [70] presentaron un análisis exhaustivo del estado del arte en segmentación de manos *in the wild* basado en aprendizaje profundo, con pruebas sobre distintas bases de datos de videos de gestos dinámicos manuales capturados *en primera persona*[34]. Si bien este último trabajo no se encuentra desarrollado en el marco del RALS, bien podría aplicarse para la detección de la actividad de las manos en distintos contextos.

2.7. Comentarios de fin de capítulo

En este capítulo se presentaron los principales aspectos del RALS. En particular, siguiendo las distintas etapas de la Figura 2.1, se realizó una búsqueda bibliográfica de distintas alternativas bajo el abordaje "clásico" del problema, finalizando con la presentación de soluciones basadas en aprendizaje profundo.

[34]Condición conocida comúnmente en la literatura como *egocentric vision*.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 3

Bases de datos existentes para el RALS

En este capítulo se expondrán las bases de datos de lengua de señas empleadas por la comunidad, a los fines de comprender el estado del arte actual al respecto. En esta búsqueda resultó importante notar la aplicación para la cual fue adquirida la información en cada caso. A continuación se expone una breve reseña por cada una de las bases de datos estudiadas. Por razones de tiempo se presenta aquí un conjunto acotado de las bases de datos existentes, priorizando por las bases de datos bien documentadas y de acceso público. Por una revisión más completa de las bases de datos existentes en RALS, dirigirse a [6]. Asimismo, el autor también refiere al artículo de Pisharady y Saerbeck [102], en el cual se expone una serie de bases de datos para el reconocimiento de gestos manuales, independientemente de su empleo en la lengua de señas.

Las bases de datos estudiadas se clasificaron en dos grupos. Por un lado, en la Sec. 3.1 se presentan bases de datos de imágenes estáticas, empleadas para el reconocimiento de configuraciones manuales o del deletreo manual. Por otro lado, en la Sec. 3.2 se presentan bases de datos de videos, empleadas para el reconocimiento de gestos dinámicos, involucrados en la conformación de señas aisladas, o bien de discurso continuo. En la Sec. 3.3 se presentan las características de una base de datos *benchmark* y las medidas de desempeño más frecuentes en el campo del RALS. En la Sec. 3.4 se presentarán los aspectos mínimos que deberán tomarse en cuenta para la adquisición de una base de datos propia para la implementación de un sistema de RALS uruguayo. Finalmente, en la Sec. 3.5 se exponen las conclusiones de este capítulo.

3.1. Bases de datos de gestos estáticos

3.1.1. ASL Finger Spelling Dataset

En 2011 Pugeault y cols. presentaron una base de datos denominada ASL Finger Spelling Dataset para el reconocimiento del deletreo manual [106]. La base de datos es multimodal y se encuentra compuesta por imágenes RGB y mapas de profundidad capturadas por un sensor Microsoft® Kinect™. Los registros se encuentran

organizados por sujeto, y comprenden 24 de las 26 letras del alfabeto dactilológico de Lengua de Señas Americana (ASL). Las 2 letras excluidas fueron la 'j' y la 'z' debido a que su ejecución implica movimiento, imposibilitando su caracterización completa mediante imágenes estáticas aisladas.

Las muestras de la base de datos provienen de un sistema de captura de video. Los sujetos no vestían guantes. En cinco sesiones diferentes –con condiciones de iluminación controladas– se solicitó a los sujetos hacer cada letra frente al sensor y mover la mano, manteniendo la configuración manual, a los fines de tener distintas perspectivas y fondos. De esta manera se registraron aproximadamente 500 muestras por seña, provenientes de 5 sujetos (no nativos de ASL) [106]. En la Figura 3.1 se observan algunas muestras de la base de datos, una por sujeto, para cada una de las letras consideradas. Observar la variabilidad anatómica de las manos, así como de tamaño y orientación en la imagen, además de la variabilidad del fondo.

Figura 3.1: ASL Finger Spelling Dataset, Muestras de Dataset A. Tomada de [106].

Posteriormente, los mismos autores presentaron otra base de datos compuesta únicamente por mapas de profundidad, con muestras provenientes de 9 sujetos y dos condiciones de iluminación y de fondo muy distintas.

Ambas bases se encuentran accesibles públicamente[1]. La primera de ellas se encuentra bajo el título de 'Dataset A: 5 users (easy)', mientras que la segunda bajo 'Dataset B: 9 users (hard)'. En cuanto a Dataset A, se encontraron algunas inconsistencias entre las descripciones del *web* y del artículo [106], tales como el número de sujetos –5 *versus* 4– y las condiciones del fondo –constante *versus* variable–.

3.1.2. NUS hand posture datasets I y II

Tanto "NUS hand posture dataset I" como "NUS hand posture dataset II" son bases de datos de imágenes de posturas manuales estáticas. Ambas fueron desarrolladas por la Universidad Nacional de Singapur, son de acceso público y gratuito para fines académicos de investigación[2].

[1] http://empslocal.ex.ac.uk/people/staff/np331/index.php?section=FingerSpellingDataset
[2] http://www.ece.nus.edu.sg/stfpage/elepv/NUS-HandSet/

Por su parte, la "NUS hand posture dataset I" fue desarrollada en el año 2010 para el reconocimiento automático de posturas manuales a partir de imágenes RGB [77]. Está compuesta por 10 posturas manuales, capturadas a distintas posiciones y tamaños en la escena, cuyo fondo es uniforme. Cuenta con 24 imágenes por cada clase, tanto en color como en escala de grises a una resolución de 160×120 píxeles. Las posturas manuales solicitadas fueron escogidas de modo de reducir la variación interclase, lo cual dificulta el problema del reconocimiento.

Figura 3.2: Muestras de las 10 clases y distintos fondos de la "NUS hand posture dataset II", grupo A. Tomada de [103].

Por otra parte, la "NUS hand posture dataset II" fue desarrollada en el año 2013 con el propósito de segmentar y reconocer posturas manuales sobre fondos complejos a partir de imágenes RGB [103]. Está compuesta por 10 clases, con varios tamaños y formas de manos. Los registros fueron obtenidos a partir de 40 sujetos, de variadas etnias contra diferentes fondos complejos y naturales. Los sujetos son de sexo masculino y femenino, de 22 a 56 años de edad. Se cuenta con 5 repeticiones de cada postura por parte de cada sujeto, dando un total de 2000 imágenes de 160×120 píxeles. Asimismo, la base de datos se encuentra subdividida en tres grupos: A, B y C. El grupo A se conforma como se explicó. En el grupo B se agrega ruido "natural" a las imágenes, tales como parte del rostro o parte del cuerpo del señante u otro individuo en el fondo. El grupo C está compuesto por imágenes de fondo únicamente. En la Figura 3.2 se observan muestras de las 10 clases que conforman el grupo A [103].

3.1.3. LSA16

En 2014 Ronchetti y cols. crearon la base de datos LSA16 con el objetivo de producir un diccionario de Lengua de Señas Argentina (LSA) y entrenar un sistema automático traductor de señas. La base de datos es de acceso público[3] y se encuentra conformada por 5 repeticiones de 16 configuraciones manuales elementales en la LSA por parte de 10 sujetos distintos, dando un total de $5 \times 16 \times 10 = 800$ imágenes. Tal como se observa en las Figuras 3.3a, 3.3b y 3.3a, los señantes vestían ropa negra y guantes de color fluorescente –rojo en mano derecha, cian en mano izquierda–, en un entorno *de interior* –o *indoor*– con fondo blanco y luz controlada [111]. El elemento sensor fue una cámara *web* RGB genérica con resolución 640×480 píxeles.

[3] http://facundoq.github.io/unlp/lsa16/index.html

(a) Registro '1_1_3.png'. (b) Registro '1_2_3.png'. (c) Registro '1_10_5.png'.

Figura 3.3: Tres muestras de imágenes crudas de la base de datos LSA16 correspondientes a la configuración manual '1' ejecutada por los sujetos '1', '2' y '10' –de izquierda a derecha–.

Las repeticiones de cada configuración manual fueron realizadas a distintas orientaciones. El nombre de archivo de cada imagen denota la configuración manual, el sujeto y la repetición. Por ejemplo, el registro '3_2_4.png' corresponde a la cuarta repetición de la configuración '3' ejecutada por el sujeto '2'.

Además de contar con las imágenes crudas de la escena completa, LSA16 posee versiones preprocesadas de las imágenes, esto es, manos segmentadas y aisladas del fondo. En la Figura 3.4 se pueden observar 16 muestras de imágenes preprocesadas, ilustrándose aquí además las 16 configuraciones manuales que componen la base de datos.

Figura 3.4: Ejemplos preprocesados de cada clase incluidos en la base de datos LSA16.

De esta manera, además de una base de datos para el reconocimiento, se dispone de material para el desarrollo de métodos de segmentación de manos, considerando las configuraciones manuales preprocesadas por los autores de LSA16 como *ground-truth* para la evaluación del desempeño.

3.1.4. RWTH-PHOENIX-Weather MS Handshapes

En 2016 Koller y cols. introdujeron RWTH-PHOENIX-Weather MS Handshapes, una base de datos para el reconocimiento automático de configuraciones manuales a través de una CNN, robusto frente a una alta similitud interclase [75].

RWTH-PHOENIX-Weather MS Handshapes fue reunida a partir de tres bases de datos: dos diccionarios *online*, uno de Lengua de Señas Danesa (DLSD) y otro de Neozelandesa (DLSNZ). La tercer base de datos fue RWTH-PHOENIX-Weather, a ser descripta posteriormente en la Sec. 3.2.4.

RWTH-PHOENIX-Weather MS Handshapes es de acceso público[4] y se compone de más de 1 millón de imágenes: 65.088 provenientes de la base de datos de DLSD de tamaño 227×227 píxeles, 153.298 de la base de datos de DLSNZ de tamaño 140×140 píxeles y 786.750 de RWTH-PHOENIX-Weather de tamaño 92×132 píxeles. Los datos provienen de 23 sujetos en total. RWTH-PHOENIX-Weather MS Handshapes cuenta con 60 configuraciones manuales, las cuales no se encuentran igualmente muestreadas. Se etiquetaron manualmente 3359 muestras de RWTH-PHOENIX-Weather y, en particular, 14 de las 60 configuraciones manuales representan aproximadamente el 90 % de las muestras etiquetadas. En la Figura 3.5 se muestran 3 muestras de 12 configuraciones manuales consideradas. Obsérvese la gran similitud entre las diferentes configuraciones manuales [75].

Figura 3.5: RWTH-PHOENIX-Weather MS Handshapes, 12 ejemplos de configuraciones manuales anotadas manualmente. Tres *frames* por clase en cada columna. Tomada de [75].

Las muestras no etiquetadas fueron igualmente empleadas por los autores para el entrenamiento de la red, bajo un esquema de aprendizaje denominado *supervisión débil* –o *weak supervision*–, el cual constituye un aspecto central del abordaje propuesto [75]. Los detalles de esta base de datos se retomarán en la Sec. 4.1.2.

3.1.5. Otras

Además de las bases de datos descriptas, durante la búsqueda se encontraron bases de datos de imágenes empleadas para abordar el problema del RALS a distintos niveles. En particular, a continuación se hace mención a dos de ellas.

Señas de LIBRAS. En 2015 Bastos y cols. presentaron una base de datos de 9600 imágenes de 40 señas de la Lengua de Señas Brasilera, entre las cuales se incluyen el alfabeto, números y algunas palabras [16].

ASLID. En 2016 Gattupalli y cols. presentaron ASLID, una base de datos de imágenes de acceso público[5], para el entrenamiento de una red profunda para el reconocimiento de posturas corporales en el marco del RALS [56].

[4] https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/.
[5] http://vlm1.uta.edu/~srujana/ASLID/ASL_Image_Dataset.html.

3.2. Bases de datos de gestos dinámicos

3.2.1. RWTH German Fingerspelling Database

En 2006 Dreuw y cols. de la Universidad de Aachenn presentaron una base de datos propia de videos de deletreo manual de Lengua de Señas Alemana [41]. La base de datos es de acceso público[6] y está compuesta por 35 gestos, incluyendo las letras de la 'A' a la 'Z', la 'SCH', las *vocales alemanas modificadas*[7] 'Ä', 'Ö', 'Ü' y los números del '1' al '5'. Se contó con la participación de 20 señantes, 2 repeticiones por gesto, una por sesión de registro, lo cual resultó en un total de 1400 registros. Los registros fueron realizados en un entorno de interior bajo condiciones de iluminación natural –luz de día– no uniforme, con perspectivas de cámara variables, y personas sin guantes ni restricciones de vestimenta. Los sujetos fueron registrados por dos cámaras simultáneas, una cámara *web* –de resolución 320×240– y una filmadora –de resolución 352×288–, ambas a 25 fps, desde puntos de vista diferentes. En la Figura 3.6 se observan tres *frames* de muestra de la base de datos, provenientes de tres sujetos. Obsérvense los distintos puntos de vista de filmación empleados.

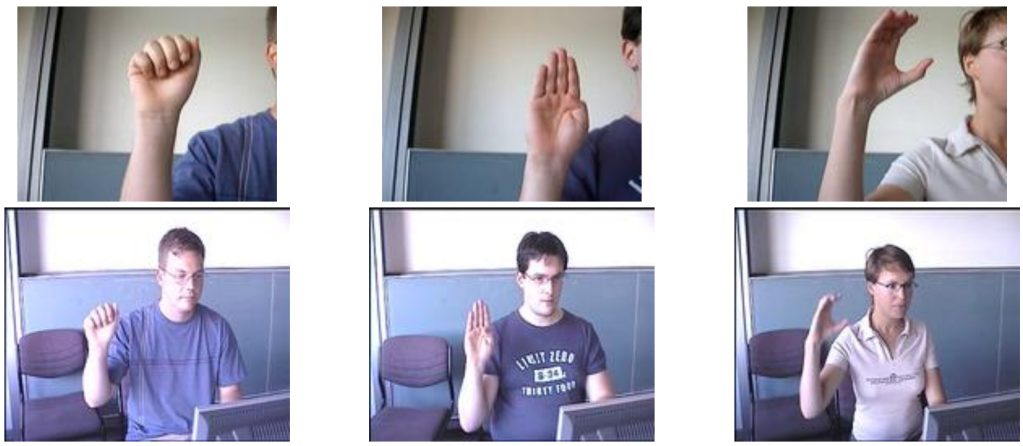


Figura 3.6: Muestras 'A', 'B' y 'C', tres sujetos, German RWTH Fingerspelling Database. Tomada del sitio *web*[6].

3.2.2. RWTH-BOSTON-50 Database

La RWTH-BOSTON-50 es una base de datos creada en 2005 para el reconocimiento de seña aislada correspondientes a la ASL. La misma no fue registrada *per se* sino creada por Zahedi y cols. [146], a partir una base de datos de oraciones de ASL publicada por el National Center for Sign Language and Gesture Resources de la Boston University[8]. En esta base de datos, las señas fueron captadas por cuatro

[6] http://www-i6.informatik.rwth-aachen.de/aslr/fingerspelling.php
[7] Traducido del alemán, *umlaut*.
[8] http://www.bu.edu/asllrp/ncslgr.html

cámaras simultáneas estacionarias, de las cuales sólo una provee una imagen color. El resto de las cámaras provee imágenes en escala de grises. Dos de estas tres cámaras se emplearon desde el frente para formar una imagen estéreo. La cámara color se ubicó entre estas dos cámaras para capturar en detalle el rostro del señante. La cámara restante proveyó un tiro de perfil de la actividad del señante. Los registros originales son de 30 fps y 312×242 píxeles.

La RWTH-BOSTON-50 es de acceso público[9] y se compone de 483 realizaciones[10] de 50 señas aisladas de ASL. Por la forma en la cual fue construida, no todas las señas se encuentran igualmente muestreadas. Por ejemplo, se tienen 31 repeticiones de la seña 'buy', mientras que de la seña 'write' se tienen sólo 2. Los videos fueron registrados a 30 fps, con una resolución de 195×165 píxeles –se hizo un *cropping* de las escenas originales, conservando únicamente las vecindades del señante–. Los datos provienen de 3 sujetos –1 hombre y 2 mujeres–, sin uso de guantes ni restricciones de vestimenta. La base de datos original empleó 4 cámaras simultáneas, de las cuales se retuvieron sólo 2, una de frente y otra de perfil. En la Figura 3.7 puede observarse 1 *frame* de muestra, por cámara, por sujeto participante. En relación al uso de la información de ambas cámaras, los autores afirman que resulta conveniente una ponderación dominante de las características provenientes de la cámara de perfil –0.62 *vs* 0.38– [146].

Figura 3.7: RWTH-BOSTON-50, muestras de los señantes y los puntos de vista conservados. Tomada del sitio *web*[9].

Asimismo, el grupo de RALS de la RWTH Aachen University[11] –del cual Dreuw es miembro– ha registrado otras bases de datos, tanto para el reconocimiento de seña aislada como de discurso continuo. A continuación se describen dos de ellas.

[9] http://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-50.php
[10] Traducido del inglés, *utterances*.
[11] http://www-i6.informatik.rwth-aachen.de/aslr/index.php

3.2.3. RWTH-BOSTON-104 Database

En 2007 Dreuw y cols. presentaron la base de datos RWTH-BOSTON-104, la cual fue creada para el reconocimiento de oraciones de ASL [42, 46]. La misma fue obtenida de forma similar a RWTH-BOSTON-50, proveniente de la misma base de datos de origen. La RWTH-BOSTON-104 es de acceso público[12] y consiste de 201 oraciones y un vocabulario de 104 señas en videos de 312×242 de resolución a 30 fps. Los datos provienen de 3 señantes –2 mujeres y 1 hombre–. No está reportado claramente cuál de las 4 cámaras de la base de datos de origen fue conservada.

Figura 3.8: RWTH-BOSTON-104, muestras de los señantes. Tomada del sitio *web*[12].

El *corpus* se encuentra dividido en 161 oraciones para el entrenamiento y 40 para el testeo. Debido a que algunas palabras resultaron de ocurrencia en la base de datos, los autores tuvieron problemas para llevar a cabo su reconocimiento [46].

3.2.4. RWTH-PHOENIX-Weather

En 2012 Forster y cols. presentaron RWTH-PHOENIX-Weather, una base de datos creada para el RALS y la traducción [54]. En particular, RWTH-PHOENIX-Weather se compone de registros de discurso continuo en DGS sobre el pronóstico del tiempo por parte de varios intérpretes. RWTH-PHOENIX-Weather fue registrada a partir de la grabación durante 3 años de una serie de noticieros emitidos por el canal televisivo alemán 'Phoenix', de allí su nombre. RWTH-PHOENIX-Weather es una base de acceso público[13] y está compuesta por 190 videos con una resolución de 210×260 píxeles a 25 fps. Los registros fueron realizados bajo condiciones controladas de un estudio de TV. Los intérpretes –6 mujeres y 1 hombre– eran todos de mano dominante derecha y vestían ropa oscura, sobre un fondo gris artificial con transición de color. En la Figura 3.9 se puede ver una imagen de ejemplo por cada intérprete de la base de datos, así como la distribución de registros por sujeto.

Si bien se trata de una base de datos compuesta por registros realistas, es preciso mencionar algunas particularidades. En primer lugar, debido a la temática del discurso, el léxico empleado es limitado. En segundo lugar, los señantes son intérpretes oyentes traduciendo en tiempo real, lo cual implica dos aspectos. Por un

[12] http://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-104.php
[13] https://www-i6.informatik.rwth-aachen.de/~forster/database-rwth-phoenix.php.

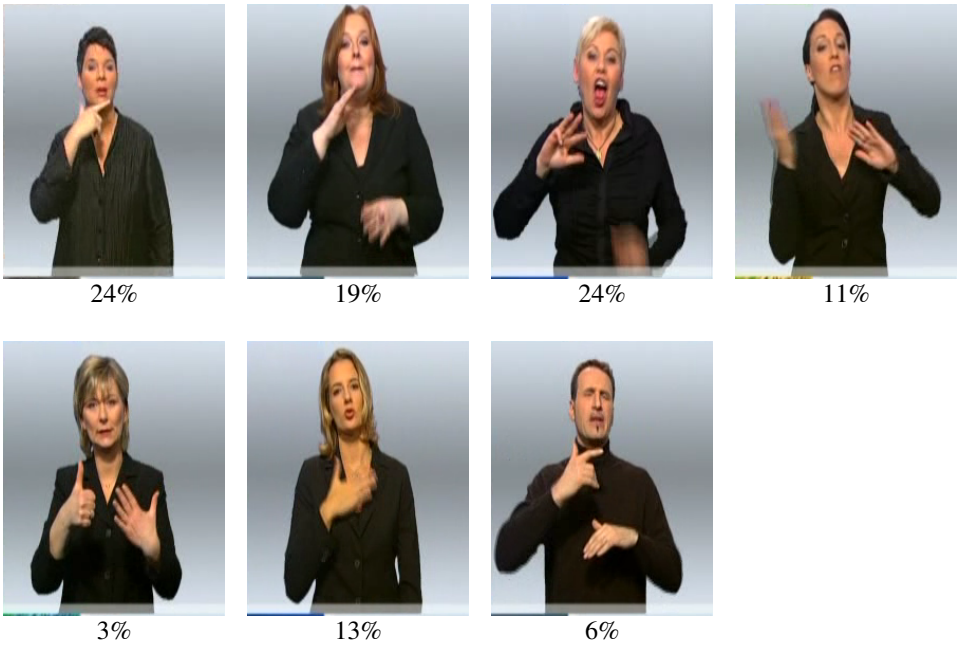


Figura 3.9: RWTH-PHOENIX-Weather, imágenes de ejemplo y distribución de los datos según intérprete. Tomada de [54].

lado, la estructura gramatical de la DGS no está completamente conservada. Por otro lado, algunas señas no son ejecutadas completamente debido a la traducción en tiempo real [54].

Cada uno de los videos del *corpus* fue etiquetado mediante el *software* ELAN[14], de uso específico para anotaciones en audio y video [54]. En particular, cada video fue segmentado en tiempo a nivel oración y a nivel palabra, por parte de un sujeto sordo experto. Además, se agregó a cada segmento la palabra y oración correspondiente en alemán escrito [54]. RWTH-PHOENIX-Weather cuenta además con el etiquetado a nivel imagen de 266 *señas*. Para ello se aislaron muestras de video y en cada *frame* se etiquetaron rasgos manuales y faciales. En particular, en cada *frame* se etiquetó el centroide de las manos y la punta de la nariz. Luego, sobre un subconjunto de 369 imágenes el Institute of Interactive and Itelligent Systems de la University de Innsbruck anotó 38 puntos prominentes sobre el rostro de los 7 intérpretes, cubriendo una gran variedad de expresiones y orientaciones faciales [54].

En la Figura 3.10 se observan algunos ejemplos de estas anotaciones; al centro, el *tracking* de las manos y la punta de la nariz; y, a los lados, el etiquetado sobre el rostro de dos intérpretes.

[14]https://tla.mpi.nl/tools/tla-tools/elan/.

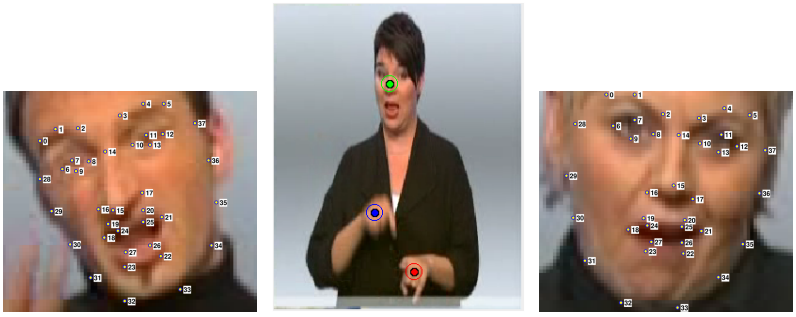


Figura 3.10: RWTH-PHOENIX-Weather, visualización de anotaciones de *tracking* (centro) y etiquetado de rostro (a los lados). Tomada de [54].

3.2.5. SIGNUM

La base de datos SIGNUM[15] fue creada en 2007 por Von Agris y cols. para el desarrollo de un sistema de reconocimiento de lengua de señas basado en video robusto frente a la variabilidad intersujeto [135]. SIGNUM está compuesta tanto por señas aisladas como discurso continuo, provenientes de 25 señantes nativos de diferente sexo y edad. El léxico registrado se constituye de 450 señas básicas[16] de la Lengua de Señas Alemana (DGS). En base a estas señas, se construyeron y registraron 780 oraciones, las cuales comprenden desde 2 a 11 señas cada una. La selección de señas y oraciones registradas permite indagar sobre aspectos de importancia en la DGS, tales como señas compuestas y la influencia de los rasgos faciales en la determinación de las señas [135]. Uno de los sujetos participantes se escogió como *señante de referencia*, del cual se registraron 3 repeticiones por seña frente a 1 repetición para el resto. En total, SIGNUM se compone de 33.210 registros, 12.150 de señas aisladas y 21.060 de oraciones. A su vez, a los fines de evaluar el desempeño frente a distintos tamaños de vocabulario, SIGNUM se encuentra dividida en tres *subcorpus*, con 150, 300 y 450 señas, respectivamente [135].

Los registros de video fueron realizados con una cámara industrial –en particular, la AVT Marlin F-046C– con una resolución de 776×578 píxeles a 30 fps[17]. La cámara AVT Marlin F-046C permitió llevar a cabo los registros de forma prácticamente automática a partir de la implementación de un *software* de control *ad hoc*. En la Figura 3.11 pueden observarse tres *frames* de muestra de la base de datos.

Los registros se llevaron a cabo bajo condiciones de laboratorio, es decir, fondo monocromático azul e iluminación controlada. En particular, la iluminación se realizó con luz proveniente de 6 lámparas, difundida a los fines de reducir la presencia de penumbras. Cada señante vestía ropa oscura manga larga y fue ins-

[15]Por sus siglas en inglés, *Signer-Independent Continuous Sign Language Recognition for Large Vocabulary Using Subunit Models*.

[16]Los autores refieren como 'básica' a toda seña (1) frecuente en la cotidianeidad, y (2) indivisible en señas más pequeñas [135].

[17]Por más detalles técnicos sobre la adquisición de datos –lentes, iluminación y montaje–, dirigirse a http://www.bas.uni-muenchen.de/Bas/SIGNUM/.

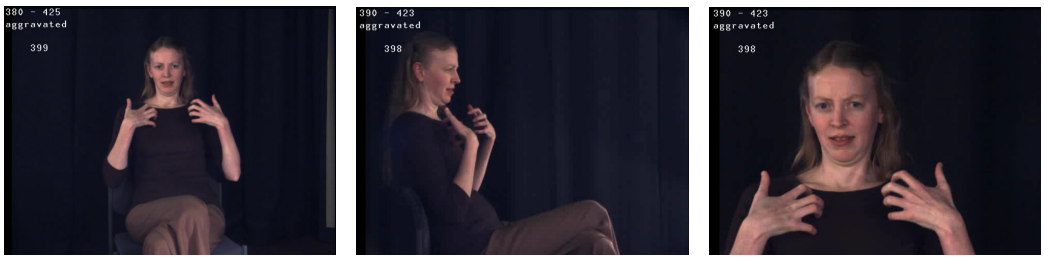
Figura 3.11: SIGNUM, *frames* de ejemplo tomados de tres señantes nativos de diferente sexo y edad. Tomada de [135].

truido para mover sus manos para permanecer en el mismo sitio y realizar las señas desde una posición de las manos de reposo y volver a ésta –manos a los lados de la cadera–. Las manos permanecen visibles durante toda la ejecución, y parten y finalizan en la misma posición, lo cual simplifica la identificación y el *tracking*.

SIGNUM es una base de acceso público. No obstante, su elevado peso –aproximadamente 1 TByte– no permite su descarga via *web*, con lo cual debe solicitarse su envío en un medio de almacenamiento masivo directamente a los autores.

3.2.6. ASLLVD

En 2012 Neidle y cols. de la Boston University propusieron una base de datos denominada ASLLVD[18] para el desarrollo de tecnologías para la búsqueda de señas a partir de video [93]. La misma es de acceso público[19] y está compuesta por casi 9800 muestras. El *corpus* se encuentra conformado por más de 3300 señas aisladas de ASL por parte de 1 a 6 señantes nativos. Cada registro se compone por 4 tomas de video simultáneas desde diferentes perspectivas: una toma lateral, una toma cercana a la región de la cabeza y dos tomas frontales con distinta resolución. En la Figura 3.12 pueden observarse tres *frames* capturados simultáneamente.



(a) Toma de frente. (b) Toma de perfil. (c) Toma del rostro.

Figura 3.12: ASLLVD, muestras de capturas simultáneas de la base de datos. Tomadas de sitio *web*[19].

[18]Por sus siglas en inglés, *American Sign Language Lexicon Video Dataset*.
[19]http://www.bu.edu/av/asllrp/dai-asllvd.html

ASLLVD cuenta con múltiples anotaciones lingüísticas, incluyendo tiempos de comienzo y fin de cada seña, etiquetas de configuración manual inicial y final para cada mano en cada seña, como así también clasificaciones según la morfología y articulación de la seña. Incluso, para las señas compuestas –749 muestras– esta base de datos posee anotaciones a nivel de morfema, es decir, de las sub-unidades componentes. Durante el registro, las señas fueron solicitadas haciendo uso de un diccionario previo y en algunos casos ocurrió que la misma seña fue ejecutada de formas distintas según el señante. En cuanto al RALS, la base de datos posee etiquetas para identificar correctamente las señas y sus variantes, los videos se encuentran en formato crudo sin compresión y se cuenta con una rutina de calibración y un *software* para la segmentación de las manos, los brazos y el rostro por *detección de piel* –o *skin detection*– [93].

La búsqueda y descarga de los videos de ASLLVD es realizada a través de una interfaz *web* desarrollada por los autores denominada DAI [94], accesible en http://secrets.rutgers.edu/dai/queryPages/search/search.php.

A partir de ASLLVD, Neidle y cols. desarrollaron una herramienta para el análisis lingüístico y etiquetado de la Lengua de Señas denominada SignStream® [92].

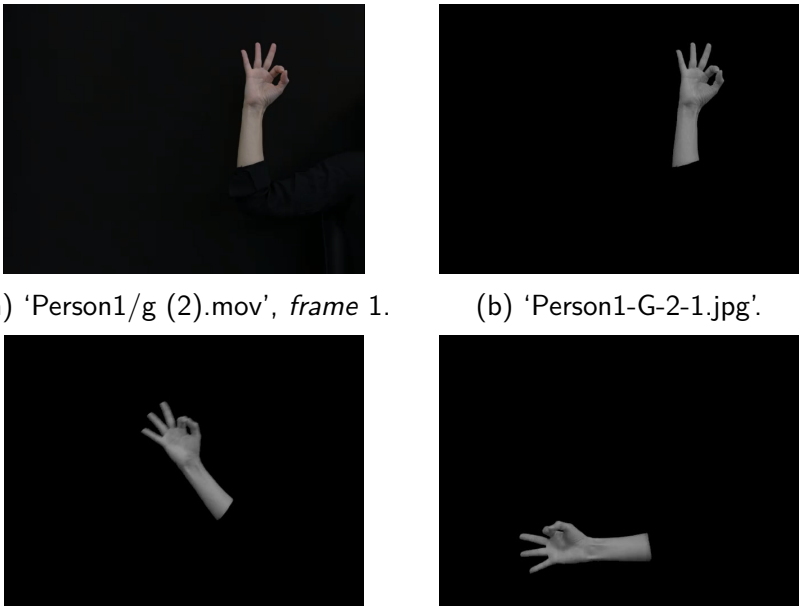
3.2.7. ISL-HS Dataset

'Irish Sign Language - Hand shape dataset' (ISL-HS) es una base de datos creada para el reconocimiento del deletreo manual de Lengua de Señas Irlandesa (ISL[20]). ISL-HS es de acceso público[21] y está compuesta por videos de las 26 letras del alfabeto dactilológico de la ISL. Todas las letras son señas unimanuales, 23 de ellas se caracterizan mediante configuraciones manuales estáticas, mientras que las 3 restantes –'J', 'X' y 'Z'– requieren de rasgos adicionales. Los datos provienen de 6 sujetos –3 hombres y 3 mujeres–, con vestimenta negra, sobre un fondo negro. Se solicitaron 3 repeticiones por letra, por sujeto, dando un total de 468 registros.

La base de datos ISL-HS se compone de videos cortos, tanto para los gestos dinámicos como para los estáticos. Cada gesto estático –configuración manual– fue realizado moviendo el antebrazo en "arco" desde la posición vertical a la posición horizontal, de modo de capturar variaciones en la orientación de cada configuración manual. Los gestos dinámicos fueron ejecutados naturalmente, esto es, sólo se capturó el movimiento implicado sin rotaciones adicionales.

La base de datos cuenta con dos directorios: los datos crudos en video y los *frames* componentes preprocesados. Los datos crudos en video se obtuvieron mediante un Apple iPhone 7, en formato '.mov', resolución 640×480, 30 fps, y RGB en 24 bits. En el directorio 'frames' se disponen las imágenes componentes de los videos. En particular, se llevaron a escala de grises y se les removió el fondo para retener sólo los píxeles de antebrazo y mano. En la Figura 3.13 se presentan 4 muestras de la segunda repetición del gesto estático 'G', por parte del sujeto 1. Obsérvense los detalles de vestimenta y de fondo en la Figura 3.13a el primer *frame* del video crudo.

[20] Por sus siglas en inglés, *Irish Sign Language*.
[21] https://github.com/marlondcu/ISL



(a) 'Person1/g (2).mov', *frame* 1. (b) 'Person1-G-2-1.jpg'.

(c) 'Person1-G-2-68.jpg'. (d) 'Person1-G-2-111.jpg'.

Figura 3.13: Muestras de la base de datos ISL-HS. Tomadas del sitio *web*[21].

A su vez, los datos se encuentran estructurados por sujeto –'Person1', 'Person2', ...–. El nombre de los archivos está dado por la letra en cuestión, seguida por un número entre paréntesis que indica la repetición. Además, sólo para el caso de los *frames*, el nombre agrega el número de *frame*, luego de la repetición y de la letra. En promedio, se tienen 2000 *frames* por letra y 9000 por sujeto.

3.2.8. LSA64

En 2016 Ronchetti y cols. [112] presentaron una base de datos de Lengua de Señas Argentina (LSA) denominada LSA64, la cual fue creada con el fin de generar un diccionario de LSA y entrenar un sistema de RALS. LSA64 es de acceso público[22] y está compuesta por 64 señas, 42 unimanuales y 22 bimanuales, ejecutadas 5 repeticiones cada una por parte de 10 sujetos no nativos, lo cual redunda en un total de 3200 videos. Las señas seleccionadas son las más empleadas del léxico de la LSA e incluyen tanto sustantivos como verbos. En la Figura 3.14 se muestran algunos *frames* de ejemplo de la base de datos.

La base de datos fue registrada en dos conjuntos. El primero fue registrado en un entorno *de exterior* –u *outdoor*– con luz natural –columna izquierda de la Figura 3.14–, mientras que el segundo lo fue en un entorno de interior con luz artificial –columnas central y derecha de la Figura 3.14–. En ambos conjuntos, los sujetos vestían guantes de color y vestimenta oscura sobre un fondo uniforme blanco.

[22]http://facundoq.github.io/unlp/lsa64/.

Figura 3.14: Muestras de *frames* crudos de 6 señas diferentes de la base de datos LSA64. Tomada de [112].

LSA64 cuenta con versiones preprocesadas de las muestras. En particular, cuenta con las manos segmentadas, a los fines de estudiar la evolución de las configuraciones manuales conforme transcurre una seña. Asimismo, LSA64 cuenta con anotaciones de la posición de las manos y la cabeza para cada *frame*. En particular, la posición de las manos se encuentra referida a la posición de la cabeza [112].

3.2.9. Otras

Además de las bases de datos descriptas, durante la búsqueda se encontraron bases de datos de video empleadas para abordar el problema del RALS a distintos niveles. En particular, se hace mención a tres de ellas.

Australian Sign Language signs (High Quality) Data Set. Propuesta por Kadous en 2002, esta base de datos es de acceso público[23] y se compone de 95 señas de la Lengua de Señas Australiana, con 27 muestras de cada una, provenientes de un señante nativo. Cada muestra está compuesta por el registro simultáneo de la actividad manual provista por un guante instrumentado (5DT) más la posición provista por marcadores magnéticos ubicados en puntos anatómicos de interés del señante, dando un total de 22 canales de información.

Libras Movement. Propuesta por Dias y cols. en 2009, Libras Movement es de acceso público[24] y contiene 15 clases con 24 muestras de cada una para el estudio del movimiento de las manos en Lengua de Señas Brasilera (LIBRAS). Se compone de 360 videos en total.

NTU RGB+D. En 2016 Shahroudy y cols. presentan esta base de datos empleada para el reconocimiento de 60 tipos de acciones humanas en un sentido amplio. La misma se compone de más de 56.000 registros de video RGB más un canal de mapa de profundidad, provenientes de 40 sujetos [117].

[23] https://archive.ics.uci.edu/ml/datasets/Australian+Sign+Language+signs%28High+Quality%29.

[24] https://archive.ics.uci.edu/ml/datasets/Libras+Movement.

3.3. *Benchmarks* y métricas de desempeño

En esta sección se expondrán dos aspectos. Por un lado, el concepto de una base de datos *benchmark* y, por otro, las métricas de uso más frecuente en la evaluación del desempeño de un sistema de RALS.

Según Cheok y cols. una base de datos *benchmark* es aquélla que permite la evaluación y comparación de soluciones de manera libre de modelos e independiente del señante [31]. Luego, una base de datos *benchmark* requiere que los registros componentes garanticen soluciones escalables y de utilidad práctica. De esta manera, una base de datos *benchmark* para el RALS debe contemplar, entre otras características, la variabilidad de las señas, la variabilidad de los señantes y la variabilidad del entorno –fondo y condiciones de iluminación–. Entre las bases de datos estudiadas en este capítulo, es posible mencionar como *benchmarks* RWTH-PHOENIX-Weather, SIGNUM, RWTH-BOSTON-50 y RWTHBOSTON-104 [44].

Para la evaluación del desempeño de un sistema de RALS se han empleado distintas métricas, las cuales dependerán del dominio en el cual se encuentre el *ground-truth*. A continuación se describen en detalle la matriz de confusión, el *accuracy*, la *word error rate*, la *position independent word error rate* y la *tracking error rate*.

3.3.1. Matriz de confusión y tasa de reconocimiento

Durante el proceso de clasificación de imágenes aisladas, por ejemplo para el reconocimiento de configuraciones manuales sobre imágenes, es frecuente computar una matriz de confusión en la cual es posible observar la proporción de cada clase bien clasificada. Esta representación del desempeño permite identificar específicamente cuáles son las clases más problemáticas y cuáles son las sustituciones realizadas durante la clasificación. A partir de ésta pueden derivarse medidas de desempeño global tales como la *tasa de reconocimiento* –o *accuracy*–, la cual se define como la proporción de clasificaciones correctas sobre el total de clasificaciones realizadas. Esta medida es la comúnmente empleada en sistemas de RALS sobre imágenes estáticas [75, 103].

3.3.2. Word Error Rate (WER)

Para la evaluación del desempeño sobre el reconocimiento de oraciones en el RALS es frecuente el uso de una medida denominada *Word Error Rate* (WER), la cual se define como [42]:

$$\text{WER} = \frac{N_I + N_S + N_D}{N_T},$$

siendo N_I, N_S y N_D el mínimo número de señas insertadas, sustituidas y *omitidas*[25] en un total de N_T señas clasificadas. Algunos autores refieren a esta medida como *Sign Error Rate* (SER) [114, 143]. También es frecuente el uso de la medida *Sign Accuracy* (SA), la cual se define como $\text{SA} = 1 - \text{WER}$ [67, 121].

[25] Traducido del inglés, *deleted*.

3.3.3. Position independent word Error Rate (PER)

Durante la evaluación de los sistemas de traducción sobre discurso continuo, algunos autores reportan una medida de desempeño denominada *Position independent word Error Rate* (PER), la cual se define como la proporción de señas mal clasificadas, independientemente de su posición en la oración [31, 84, 114].

3.3.4. Tracking Error Rate (TER)

La *Tracking Error Rate* (TER) se emplea para cuantificar el error geométrico de estimación de la posición de un objeto de interés a lo largo de un video. Dada una secuencia de imágenes $X = \{x_n\}_{n=1}^{N}$ y las posiciones de referencia de un objeto de interés $U = \{u_n\}_{n=1}^{N}$, la *tracking error rate* de las posiciones del objeto de interés detectadas $\hat{U} = \{\hat{u}_n\}_{n=1}^{N}$ se define como la proporción de *frames* para los cuales la distancia euclídea es mayor o igual a una cierta tolerancia τ [43]:

$$\text{TER} = \frac{1}{N}\sum_{n=1}^{N}\delta_\tau(u_n, \hat{u}_n), \text{ con } \delta_\tau(u,v) = \begin{cases} 0, & \text{si } |u-v| < \tau, \\ 1, & \text{en otro caso.} \end{cases}$$

En el marco del RALS la TER se ha empleado para caracterizar el error de *trackeo* de las manos. De manera lógica, aquellos *frames* en los cuales las manos no se hayan visibles no podrán emplearse para el cálculo de la TER [43].

3.4. Consideraciones para el diseño de una base de datos en LSU

Considerando lo expuesto en el Capítulo 1, para llevar a cabo el desarrollo de un sistema de reconocimiento de Lengua de Señas Uruguaya (LSU), es preciso contar con una base de datos de naturaleza local. Por razones de tiempo, la adquisición de una base de datos propia ha quedado fuera de los alcances de esta tesis de maestría. No obstante, a partir de las bases de datos estudiadas y presentadas a lo largo del presente capítulo fue posible conocer los aspectos fundamentales que deberán considerarse al momento de registrar una base de datos.

El desarrollo de una base de datos propia implica como mínimo el abordaje de las siguientes etapas:

- Elaborar un folleto informativo para cada sujeto participante, donde se explique claramente los distintos aspectos de la sesión de grabación, tales como su duración estimada, las tareas a realizar, los riesgos involucrados en el proceso y el destino que tendrán los datos recabados[26].
- Diseñar una planilla para la recolección de datos personales de interés para el RALS y problemas asociados, tales como género, edad, señante nativo/no nativo, grado de dependencia de la lengua de señas, datos antropométricos –altura y peso– y otras particularidades.

[26] http://www-i6.informatik.rwth-aachen.de/aslr/fingerspelling.php

- Elaborar un consentimiento informado, en el cual cada sujeto previamente informado adhiere con su firma a participar de la base de datos.
- Confeccionar un protocolo de adquisición y almacenamiento de los datos que garantice cierto nivel de repetibilidad en los registros y permita la sistematización en el acceso y uso de los datos almacenados;
- Convocar abiertamente a la participación de sujetos señantes, preferentemente sordos nativos de Uruguay.

De los aspectos mencionados anteriormente, a continuación se harán algunos comentarios adicionales sobre el protocolo de adquisición y almacenamiento. Desde el punto de vista metodológico, una base de datos de RALS requiere de una aplicación concreta *a priori*, en la cual se determinen los alcances de la solución tecnológica a desarrollar. Luego, será posible establecer los requerimientos de diseño de un protocolo para la adquisición y el almacenamiento de la base de datos a registrar. De manera general, y a partir de lo estudiado en este capítulo, deberán tenerse en cuenta los siguientes aspectos fundamentales:

- Vestimenta, uso/no uso de guantes de los señantes.
- Condiciones del entorno visual durante la adquisición: tipos de iluminación y fondo.
- Disposición y configuración del equipamiento de adquisición.
- Confección del *corpus* a registrar: señas y repeticiones.
- Cantidad mínima de señantes a registrar.
- Sistematización del almacenamiento de los datos.

Tomando como objetivo el caso de aplicación en Uruguay discutido en la Sec. 1.2.1 de este informe –RALS como mecanismo de búsqueda en Léxico TReLSU– se requiere una base de datos para el *reconocimiento de señas aisladas*. Un sistema de RALS realista debe ser robusto ante cambios de fondo e iluminación, ante distintos señantes y sin requerimientos específicos en cuanto a su vestimenta. Más aún, puesto que Léxico TReLSU es un diccionario de uso cotidiano, se impone como condición adicional que la solución tecnológica sea accesible, esto es, los datos de entrada deberán ser fácilmente generados por los usuarios finales. Por lo anteriormente expuesto, la base de datos deberá reunir las siguientes características:

- Un canal de registro frontal provisto por una cámara *web* RGB convencional, para el desarrollo de una solución tecnológica accesible para cualquier interesado en utilizar Léxico TReLSU.
- Registro frontal simultáneo de una cámara RGB de uso profesional, para estudiar la dependencia de las soluciones frente a la calidad de imagen.

- Registro de múltiples señantes, de ser posible en más de una sesión por señante, que permitan adquirir señas repetidas bajo distintas vestimentas.
- Registro en distintas condiciones de iluminación y fondo, para realizar el ajuste del sistema de RALS a partir de datos que cuenten con esta variabilidad.
- En primera instancia, el *corpus* estará compuesto por las 315 señas que actualmente pertenecen al Léxico TReLSU y, posteriormente, por aquéllas que eventualmente sean incorporadas.
- Almacenamiento estructurado de los datos, para un fácil acceso y segregado según seña, sujeto y sesión de grabación.

Lógicamente, cuanto mayor sea la cantidad de registros obtenidos se espera que los métodos de RALS desarrollados a partir de la base de datos sean más robustos. Para una primera etapa, se proponen los siguientes lineamientos:

- Registro de 25 señantes adultos de distintas edades y géneros, comparable a la cantidad de sujetos de la base de datos SIGNUM, *benchmark* independiente del señante –ver Sec. 3.2.5–.
- Dos sesiones de grabación. Una de las sesiones será en un entorno de interior con fondo uniforme bajo condiciones controladas de iluminación, mientras que la otra será al aire libre y con fondos de tipo "naturales".
- En la sesión de grabación de interior los sujetos emplearán vestimenta de manga larga color negra y guantes de color contrastante, un color distinto por mano. En la sesión de grabación al aire libre no habrá ningún tipo de restricciones con respecto al tipo y color de la vestimenta y los sujetos no emplearán guantes.
- En cada sesión de grabación se realizará el registro de 5 repeticiones de 50 señas a escoger entre las ≈140 señas unimanuales actualmente publicadas en Léxico TReLSU. Considerando un tiempo de ejecución de 10 segundos por seña y un margen de 20 minutos, introductorios –instrucciones y consentimiento informado– e intermedios –pausas, práctica y reajustes–, se estima una sesión de 1 hora por señante. Dentro de lo posible las 50 señas se escogerán según sean [135]: (1) frecuentes en la cotidianeidad, y (2) indivisibles en señas más pequeñas.

Llegado el caso, será necesario ahondar en mayor detalle a los fines de optimizar los tiempos –orden de las acciones–, siempre y cuando se conserve la calidad de los registros. Una vez satisfecha esta etapa se ampliará tanto el *corpus* como la cantidad de señantes y las condiciones de registro.

Para finalizar esta sección, se exponen brevemente dos ideas iniciales asociadas al desarrollo de bases de datos en LSU para el abordaje del resto de los problemas típicos del RALS presentados en la Sec. 2.1. Por un lado, para el desarrollo de un

sistema de reconocimiento de deletreo manual deberán registrarse todas las letras del alfabeto dactilológico uruguayo en formato de *video*, tomando en consideración que las letras 'G', 'H', 'J', 'Ñ', 'Q', 'X', 'Y' y 'Z' requieren del movimiento manual para ser determinadas. Por otro lado, para el desarrollo de un sistema de reconocimiento de discurso continuo, la elección de las señas y oraciones a registrar deberá ser abordada en colaboración o directamente por un lingüísta, priorizando por aquéllas de uso más frecuente. En este sentido, podrán tomarse como punto de partida los criterios de diseño de la base de datos SIGNUM, ya expuestos en la Sec. 3.2.5 de este informe.

3.5. Comentarios de fin de capítulo

En este capítulo se estudiaron las principales características de las bases de datos más importantes para el RALS. En las Tablas 3.1 y 3.2 se presenta una comparativa de las bases de datos presentadas en las Sec. 3.1 y 3.2, respectivamente. Asimismo, en la Sec. 3.3 se presentaron las principales características de una base de datos *benchmark* y las métricas más frecuentes para evaluar el desempeño de un sistema de RALS. Finalmente, a partir de la búsqueda expuesta a lo largo del presente capítulo, en la Sec. 3.4 se planteó un diseño para el registro de una base de datos en LSU para el reconocimiento de señas aisladas en el marco del diccionario Léxico TReLSU.

Nombre (Año) [Referencia bibliográfica/Fuente]	Lengua de Señas	Origen	Aplicación	# clases	# sujetos	# muestras por clase	# total de imágenes	Tipo de imágenes	Tamaño de imágenes	Contenido lingüístico de cada registro	Anotaciones
NUS Hand Posture Dataset I (2010) [77]	?	Singapur	Reconocimiento de posturas manuales	10	?	24	240	RGB	160×120	postura manual	clase
ASL Finger Spelling Dataset A (2011) [106]	norteamericana	Reino Unido	Reconocimiento de deletreo manual	24	5	≈ 500	$\approx 65000 \times 2$	RGB + mapa de profundidad	$\approx 100 \times 100$ (no uniforme)	letra del alfabeto	clase
ASL Finger Spelling Dataset B (2011) [106]	norteamericana	Reino Unido	Reconocimiento de deletreo manual	24	9	no uniforme	≈ 72500	sólo mapa de profundidad	$\approx 100 \times 100$ (no uniforme)	letra del alfabeto	clase
NUS Hand Posture Dataset II (2013) [103]	?	Singapur	Segmentación y reconocimiento de posturas manuales	10	40	200	≈ 4750	RGB	160×120 o 320×240, según subconjunto	postura manual	clase
LSA16 (2013) [111]	argentina	Argentina	Traducción automática	16	10	50	800	RGB	640×480	postura manual	clase
RWTH-PHOENIX-Weather MS Handshapes (2016) [75]	alemana, neozelandesa y danesa	Alemania	Reconocimiento de configuraciones manuales	60	23	no uniforme, ver [75]	$\approx 1 \times 10^6$	RGB	92×132, 140×140 y 227×227	configuración manual	débiles[27]

Tabla 3.1: Bases de datos de gestos estáticos relevadas.

[27] Traducido del inglés, *weakly labeled*. En este caso se refiere a una condición en la cual no todas las imágenes de una secuencia cuentan con su etiqueta correspondiente. Para más detalle, dirigirse a [75].

Nombre (Año) [Referencia bibliográfica/Fuente]	Lengua de Señas	Origen	Aplicación	# clases	# sujetos	# muestras por clase	# total de videos	Resolución de los registros	Contenido lingüístico de cada registro	Anotaciones
RWTH-BOSTON-50 Database (2005) [146]	norteamericana	Alemania	Reconocimiento de seña aislada	50	3	no uniforme	483×2	2 capturas simultáneas a 30 fps, 195×165	palabra	clase
RWTH German Fingerspelling Database (2006) [41]	alemana	Alemania	Reconocimiento de deletreo manual	35	20	2	1400×2	2 capturas simultáneas a 25 fps, 320×240 y 352×288	letras, números y *umlauts*	clase
RWTH-BOSTON-104 Database (2007) [42, 46]	norteamericana	Alemania	Reconocimiento de oraciones	?	3	?	201 (161 para entrenamiento más 40 para testeo)	30 fps, 312×242	oración	oración
RWTH-PHOENIX-Weather (2012) [54]	alemana	Alemania	Traducción automática	?	7	no uniforme	190	25 fps, 210×260	oración	clase, manos y rostro[28]
SIGNUM (2007) [135]	alemana	Alemania	Reconocimiento independiente de sujeto	450 señas y 780 oraciones	25	1 o 3 (3 para señante de referencia)	33210 (12160 señas aisladas y 21060 oraciones)	30 fps, 776×578	palabra aislada y oración	palabra aislada y oración, caracterización antropométrica y lingüística de cada señante
ASLLVD (2012) [93]	norteamericana	EEUU	Búsqueda de señas a partir de video	≈ 3300	1 a 6	no uniforme	9800×4	4 capturas simultáneas, 3 a 60 fps y 640×480 más 1 a 30 fps y 1600×1200	palabra aislada	múltiples[29]
ISL-HS Dataset (2017)[30]	irlandesa	Irlanda	Reconocimiento del deletreo manual	26	6	3	468	30 fps, 640×480	letra del alfabeto	clase
LSA64 (2016) [112]	argentina	Argentina	Reconocimiento automático	64	10	5	3200	60 fps, 1920×1080	palabra	clase más posición de la cabeza y las manos

Tabla 3.2: Bases de datos de gestos dinámicos relevadas.

[28] Posición de la nariz y centroide de las manos, más 38 *keypoints* marcados sobre el rostro [54].

[29] Múltiples anotaciones lingüísticas: tiempos de inicio y fin de la seña, configuraciones manuales inicial y final, morfemas componentes y variantes de una seña [93].

[30] Repositorio GitHub: https://github.com/marlondcu/ISL.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 4

Descripción de un sistema de RALS

En este capítulo se describe un sistema de RALS y la metodología empleada para llevar a cabo su evaluación bajo distintas condiciones. En la Sec. 4.1 se presenta el sistema bajo estudio y se exponen sus principales características; comenzando con una descripción general y siguiendo con los detalles de implementación por parte de sus autores. Posteriormente, en la Sec. 4.2 se presentan las bases de datos empleadas para la evaluación del sistema bajo estudio en este trabajo de maestría. Luego, en la Sec. 4.3 se presenta el preprocesamiento de las imágenes de acuerdo a los requerimientos del sistema escogido. Por último, en la Sec. 4.1.3 se comentan las variantes de aprendizaje por transferencia exploradas durante este trabajo.

4.1. Sistema bajo estudio: Deep Hand

En esta sección se describe el sistema de RALS estudiado durante esta tesis de maestría, introducido por Koller y cols. en el artículo "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled", en el cual se implementa y entrena una CNN para el reconocimiento de 60 *configuraciones manuales* propias de la DGS [75]. En adelante, dicho sistema será referido como Deep Hand. A continuación se comenzará su descripción exponiendo la motivación de sus autores. Luego, se presentarán los detalles de implementación de este sistema por parte de sus autores, tales como la arquitectura interna y las bases de datos y estrategias empleadas para su entrenamiento.

4.1.1. Motivación

Desde hace unos años las CNNs han mostrado muy buen desempeño en el reconocimiento de configuraciones manuales en el contexto de RALS [75]. Sin embargo, este abordaje presenta dos limitaciones serias: (1) se requiere una gran cantidad de datos para su entrenamiento, y (2) la clasificación de configuraciones manuales no es lo suficientemente exacta para permitir el RALS [75].

Tal como se presentó en las Sec. 3.1 y Sec. 3.2, las investigaciones recientes sobre la lengua de señas *per se* y el RALS han dado lugar a diversas bases de

datos, en algunos casos con disponibilidad de anotaciones de las configuraciones manuales involucradas. Este hecho motivó a Koller y cols. a desarrollar un sistema de RALS basado en los siguientes objetivos [75]:

- Hacer uso de una CNN como sistema de extracción de características y clasificación.
- Estudiar la capacidad del sistema para discriminar 60 clases de configuraciones manuales con *gran similitud* y provenientes de distintas lenguas de señas.
- Entrenar el sistema mediante *aprendizaje débilmente supervisado*[1], a los fines de poder emplear bases de datos "débilmente" etiquetadas.

A continuación se explicará con mayor detalle cada uno de los aspectos involucrados en el diseño y la implementación de Deep Hand por parte de sus autores.

4.1.2. Detalles de implementación

Arquitectura de la CNN empleada

En cuanto a la arquitectura de Deep Hand, Koller y cols. hicieron uso una versión ligeramente modificada de la red GoogLeNet, la cual se detallará en el apartado "Modificaciones de GoogLeNet y ajuste fino".

Por su parte, GoogLeNet fue introducida por Szegedy y cols. [125] y es actualmente conocida como Inception-v1, debido a la existencia de versiones ulteriores. En la Figura 4.1 se presenta un esquema de la arquitectura de la red GoogLeNet.

Como puede observarse GoogLeNet es una red de 22 capas, conformada en su mayoría por una serie de módulos característicos denominados 'inception', los cuales serán explicados en el párrafo siguiente. Las capas *totalmente conectadas*[2] y todas las capas de convolución, incluso aquéllas internas a los módulos 'inception', hacen uso de ReLU como función no lineal de activación. El *campo receptivo*[3] de GoogLeNet es de 224×224 píxeles, sobre una imagen RGB a la cual se le ha sustraído previamente la media [125].

En cuanto a los módulos 'inception' de la red, vale destacar que son empleados para reducir la cantidad de parámetros del sistema. Como se muestra en la Figura 4.2, dentro de cada módulo se realizan 4 operaciones en paralelo: 3 convoluciones con distinto tamaño de filtro más un *max pooling* de 3×3. Luego, a partir de la concatenación de los 4 resultados paralelos, se compone un volumen único de salida de cada módulo, el cual constituye a su vez la entrada del módulo subsecuente. Un problema que surge directamente de esta topología es que, incluso para un número pequeño de filtros de convolución, la concatenación da lugar a volúmenes de cantidad de canales progresivamente creciente, los cuales se vuelven intratables desde el punto de vista práctico [125].

[1] Traducido del inglés, *weakly supervised learning*.
[2] Traducido del inglés, *fully connected*.
[3] Traducido del inglés, *receptive field*.

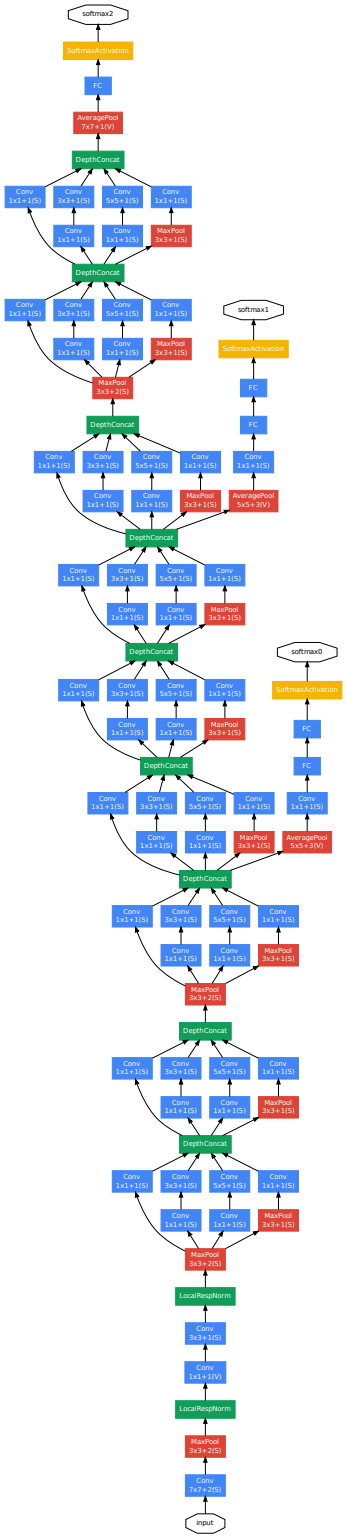


Figura 4.1: Arquitectura de la red GoogLeNet. Tomada de [125].

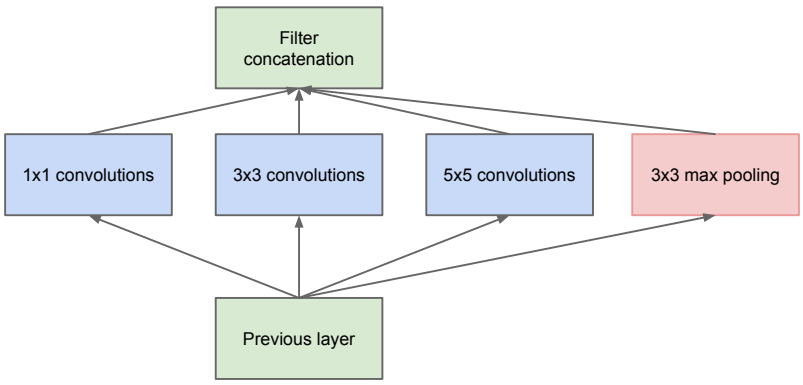


Figura 4.2: Versión original del módulo 'inception'. Tomada de [125].

Una estrategia empleada para abordar este problema es efectuar la operación de 'convolución de 1×1', la cual no es una convolución en el sentido usual, sino una reducción de la dimensión en el sentido de los canales [125]. Bajo el uso de esta operación, en la Figura 4.3 se presenta la versión del módulo finalmente implementada, la cual introduce en cada rama una operación de convolución de 1×1 para reducir la dimensionalidad de la forma ya comentada [125].

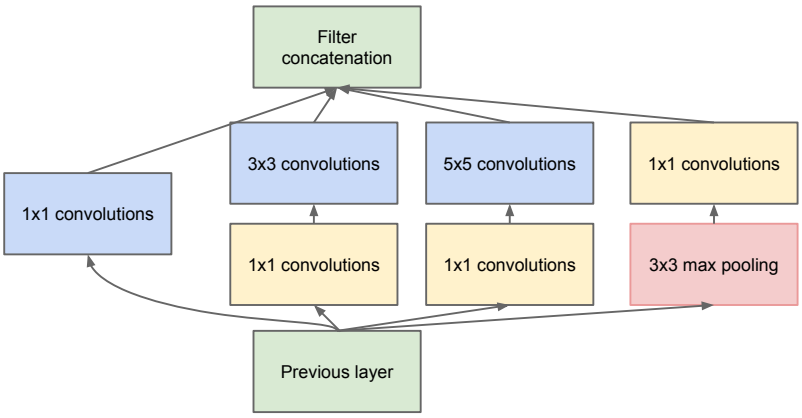


Figura 4.3: Módulo 'inception' con reducción de dimensionalidad. Tomada de [125].

Datos para el entrenamiento

En cuanto a los datos empleados para el entrenamiento de Deep Hand, los autores introdujeron una base de datos denominada RWTH-PHOENIX-Weather MS Handshapes –en adelante, denominada PhWMSHS–, de acceso público[4], compuesta por más de 1 millón de imágenes previamente segmentadas en torno a la mano derecha, provenientes de 23 sujetos en total y de 3 bases de datos de distintas lenguas de señas, a saber [75]:

[4]https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/.

- 786.750 *frames* de tamaño 92×132 píxeles, provenientes de la base de datos RWTH-PHOENIX-Weather, ya descripta en la Sec. 3.2.4,
- 153.298 *frames* de tamaño 140×140 píxeles, provenientes del diccionario *online* de Lengua de Señas Neozelandesa (DLSNZ) [37],
- 65.088 *frames* de tamaño 227×227 píxeles, provenientes del diccionario *online* de Lengua de Señas Danesa (DLSD) [66].

DLSD y DLSNZ están compuestas por registros de seña aislada, mientras que RWTH-PHOENIX-Weather por registros de discurso continuo en DGS. En la Figura 4.4 se ilustran tres secuencias de muestras, una por cada base de datos empleada; de arriba hacia abajo, DLSD, DLSNZ y RWTH-PHOENIX-Weather.

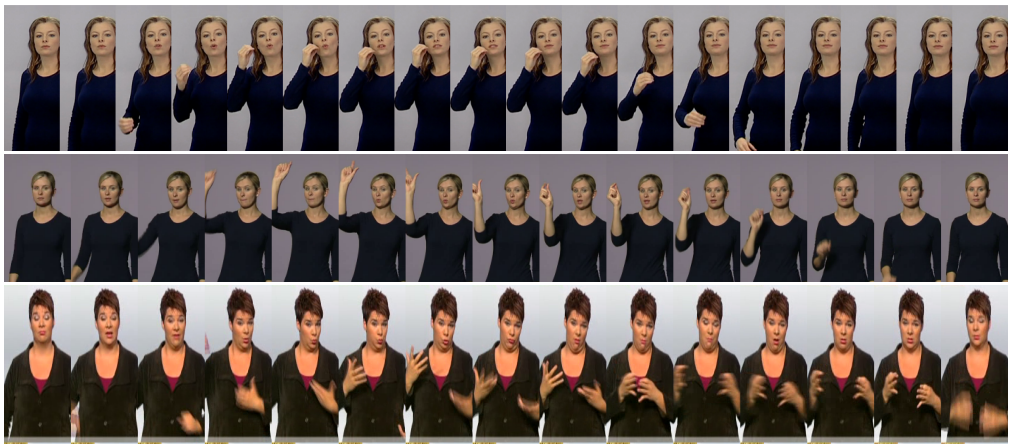


Figura 4.4: Bases de datos para entrenamiento. De arriba hacia abajo: DLSD, DLSNZ y RWTH-PHOENIX-Weather. Tomada de [75].

Vale comentar que conforme transcurre una seña a lo largo de un registro de video, las manos pueden desplazarse y sufrir cambios de orientación e incluso cambiar la configuración manual, ocasionando que algunas muestras posean *desenfoque de movimiento* –o *blurring*– en las manos. Este es el caso del final de la secuencia central y varios *frames* de la secuencia inferior de la Figura 4.4 [75]. Considerando las configuraciones manuales como etiquetas para los datos, este hecho implica un proceso de etiquetado a nivel de *frame* muy ruidoso y ambiguo. En este sentido, las bases DLSD y DLSNZ sólo poseen una o dos anotaciones por video, ya sea indicando las configuraciones manuales inicial y final, o bien una configuración manual intermedia de particular importancia lingüística [75].

Por su parte, las muestras de la base de datos RWTH-PHOENIX-Weather no cuentan con etiquetas de las configuraciones manuales. Para ello, los autores emplearon un recurso *online* de entrada abierta llamado SignWriting, el cual constituye un traductor de la lengua escrita a lengua pictórica. Esta última posee información sobre las configuraciones manuales y por tanto, Koller y cols. las emplearon para el etiquetado de esta base. En 2016 el léxico alemán de SignWriting poseía aproximadamente 25.000 entradas [75].

En base a lo expuesto, se observa que PhWMSHS no posee las etiquetas de configuración manual a nivel de cada muestra de imagen, sino que a lo sumo cuenta con una o dos etiquetas a nivel de secuencia de imágenes. Este hecho motivó a Koller y cols. a recurrir a un método de aprendizaje débilmente supervisado, el cual será comentado en el siguiente apartado.

Aprendizaje débilmente supervisado

El aprendizaje débilmente supervisado se implementó mediante un algoritmo denominado *expectation maximization*, que busca obtener etiquetas correctas a nivel de *frame* a partir de etiquetas ruidosas a nivel de video, modelando la secuencia de etiquetas como emisiones de un HMM. Para ello se ejecutan dos etapas de modo iterativo: (i) actualizar la asignación de clases a la secuencia de imágenes de entrada y (ii) actualizar los parámetros de la red en función de las nuevas etiquetas. Por más detalles, dirigirse a [75].

Preprocesamiento de los datos

Cada imagen de entrada a Deep Hand fue conformada por un *patch* en torno a una de las manos, con una extensión de 2 a 3 veces el tamaño de la mano. Asimismo a cada *patch* se le quitó la *media por píxel* de forma previa a su ingreso a Deep Hand [75].

Modificaciones de GoogLeNet y ajuste fino

En particular, Koller y cols. partieron de la red GoogLeNet, pre-entrenada para la discriminación de 1000 clases en el concurso ILSVRC 2014 [75]. Luego, para adaptar la red al reconocimiento de 60 configuraciones manuales distintas, los autores realizaron un *reemplazo* y entrenamiento de las capas totalmente conectadas previas a cada salida, cada una con 61 neuronas, 60 clases más una clase *basura* –o *garbage*– y cuyos pesos fueron inicializados en cero. Para llevar a cabo el ajuste fino del modelo, se empleó el método de descenso por gradiente con la 'entropía cruzada basada en *softmax*' como función de costo E [75]:

$$E = -\frac{1}{N}\sum_{n=1}^{N} \log p(k|x_n),$$

siendo x_n cada imagen de entrada preprocesada según se explicó.

4.1.3. Reproducción del sistema

A partir de una búsqueda en repositorios, se encontró que Necati Camgoz ofrece una implementación de Deep Hand en `TensorFlow`, de acceso libre en su repositorio de GitHub: https://github.com/neccam/TF-DeepHand. Esta implementación fue testeada sobre una base de datos de prueba de Deep Hand, descripta luego en la Sec. 4.2.1, brindando un *accuracy* del $\approx 85\,\%$.

Debido a que `TensorFlow` no es un entorno de desarrollo lo suficientemente amigable para los usuarios de `Python`, durante este trabajo de maestría se decidió traducir la implementación de Camgoz a `PyTorch`. El correcto funcionamiento de Deep Hand en `PyTorch` se verificó al obtener un desempeño muy similar sobre la misma base de datos de prueba de Deep Hand.

`PyTorch` es un paquete de cómputo basado en `Python`, cuyo objetivo es reemplazar el paquete `NumPy` de modo de optimizar el uso de GPUs[5]. `PyTorch` constituye una plataforma para la investigación del aprendizaje profundo de manera flexible y amigable, con una comunidad muy activa y gran cantidad de foros y documentación en línea. Asimismo, `PyTorch` posee un paquete muy útil para el desarrollo de herramientas de visión por computadora denominado `torchvision`, el cual recopila las bases de datos, implementaciones de CNNs pre-entrenadas y transformaciones de imagénes más difundidas en el área[6].

4.2. Bases de datos empleadas

En esta sección se presentan las 4 bases de datos empleadas para la evaluación del sistema Deep Hand en el marco de este trabajo de maestría. En primer lugar, se describe la base de datos de prueba de Deep Hand. Luego, se introduce una base de datos de configuraciones manuales propias de la Lengua de Señas Uruguaya denominada 'TReLSU-HS', fruto del trabajo de esta tesis. Por último, se presentan dos bases de datos de deletreo manual, tanto en DGS como en ASL. En particular sobre las tres últimas, se discute el proceso de conformado de las bases y los criterios seguidos para el etiquetado de los datos durante la realización de este trabajo.

4.2.1. Base de datos de prueba de Deep Hand (DH_Test)

Junto a la base de datos descripta en la Sec. 3.1.4, Koller y cols. proveen una base de datos de prueba de Deep Hand[7], compuesta por 3359 muestras etiquetadas *manualmente* [75]. En la Figura 4.5 se presentan algunas muestras de esta base de datos.

Cabe remarcar que las imágenes ya se encuentran segmentadas en torno a la mano, con un tamaño de 132×92 píxeles, con una relación de aspecto de $\frac{23}{33}$, restando aún la remoción de media por píxel de acuerdo con el apartado "Preprocesamiento de los datos" de la Sec. 4.1.2. Se destacan algunas particularidades del conjunto de prueba: imágenes con y sin rostro –fila superior *versus* fila inferior de la Figura 4.5–, presencia dos manos con la misma configuración manual –fila inferior a la derecha de la Figura 4.5–. En aquellos casos que la mano se encontraba próxima al borde de la escena completa –fila inferior a la izquierda de la Figura 4.5– los autores completaron el *patch* segmentado mediante *zero-padding*. En adelante, se hará referencia a esta base de datos mediante 'DH_Test'.

[5] https://pytorch.org/tutorials/beginner/blitz/tensor_tutorial.html.
[6] https://pytorch.org/docs/stable/torchvision/.
[7] https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/.

Figura 4.5: Muestras de la base de datos de prueba de Deep Hand. Reproducida de [4].

4.2.2. TReLSU-HS

Para la conformación de esta base de datos propia, se partió de los registros de Léxico TReLSU. Por su parte, Léxico TReLSU fue descargada mediante la aplicación `wget`, contando con un total de 315 señas, tanto unimanuales como bimanuales. Tal como se explicó en la Sec. 1.2.1, las señas de Léxico TReLSU fueron indexadas por sus autores a partir de una codificación *ad hoc* [100] según la configuración inicial y final de las manos. Luego, el nombre de cada video está compuesto por 2 configuraciones en el caso de una seña unimanual, y 4 en el caso de seña bimanual.

Etiquetado manual de los datos. Cada video de Léxico TReLSU posee una o dos etiquetas a nivel de secuencia. A diferencia de Koller y cols. [75], en este trabajo se llevó a cabo el etiquetado *manual* de los datos a nivel de *frame*. Para ello, (1) se seleccionó el subconjunto de señas *unimanuales*; y (2) se anotaron uno a uno los límites temporales de las configuraciones manuales involucradas en cada video, para lo cual se hizo uso del *software* ELAN 5.2[8] en modo 'segmentación'. En particular, las anotaciones realizadas se volcaron a un archivo de texto con el siguiente formato en cada línea: "nombre de video, configuración manual, tiempo de inicio, tiempo de fin, configuración manual, tiempo de inicio, tiempo de fin, ...". Mediante este proceso se segmentaron temporalmente las configuraciones manuales en las 133 señas unimanuales seleccionadas.

Posteriormente, haciendo uso conjunto de las herramientas `ffmpeg` y `ffprobe` desde `Python` se extrajeron y etiquetaron cada uno de los *frames* comprendidos en los períodos identificados. Cada imagen fue etiquetada forma consistente con las clases de salida de Deep Hand –ver Tabla A.1 del Anexo A–. Por completitud, en la Tabla C.1 del Anexo C se explicita la equivalencia de clases utilizada. Durante este proceso se descartaron 5 videos por encontrarse adquiridos en una perspectiva no frontal o por no encontrarse correspondencia con ninguna etiqueta de Deep Hand.

[8]ELAN 5.2 es un *software* de acceso libre ampliamente difundido entre los lingüístas para realizar anotaciones sobre datos audiovisuales [45].

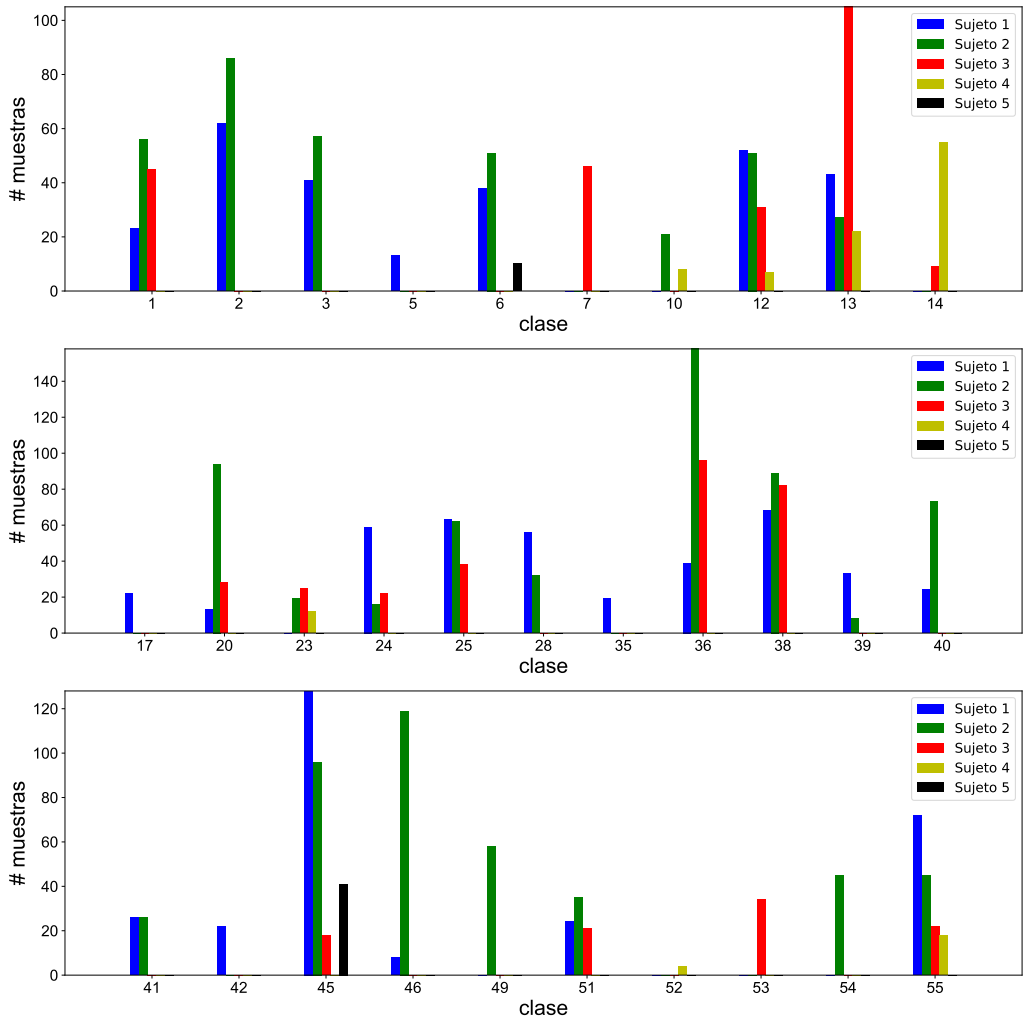


Figura 4.6: TReLSU-HS: Distribución de las 31 clases en los 5 sujetos etiquetados.

En adelante, se denominará 'TReLSU-HS' a la base de datos que resultó de este trabajo. TReLSU-HS se encuentra compuesta por 3071 *frames* etiquetados de acuerdo a las clases propuestas para Deep Hand, contando con muestras de 31 clases distintas, provenientes de 5 sujetos –3 hombres y 2 mujeres– en un entorno de condiciones controladas –fondo celeste, señantes de ropa negra e iluminación artificial–. En la Figura 4.6 se presenta la distribución de cantidad de muestras por clase para cada uno de los sujetos etiquetados.

Debe notarse que TReLSU-HS comprende sólo 31 de las 60 clases propuestas para Deep Hand. Por la manera en que fue obtenida, TReLSU-HS es una base de datos *no balanceada*, tanto en la cantidad de muestras por clase como en las clases muestreadas por sujeto. No obstante, TReLSU-HS es una base de datos útil a los fines de obtener nociones del desempeño de Deep Hand sobre configuraciones manuales de la Lengua de Señas Uruguaya.

4.2.3. RWTH German Fingerspelling Database (DGS-FS)

RWTH German Fingerspelling Database fue descripta en la Sec. 3.2.1. En particular, durante este trabajo se emplearon sólo los registros provistos por la 'cámara 2' –ver muestras en el inferior de la Figura 3.6–[9]. Según lo reportado, RWTH German Fingerspelling Database se encuentra compuesta por 2 realizaciones –1 por sesión de grabación– de 35 señas por parte de 20 sujetos [41]. No obstante, se encontró que la base de datos posee registros provenientes de 26 sujetos, siendo necesario aclarar que los registros de los sujetos 21 a 26 cuentan con una sola realización y que los registros de los sujetos 21, 22 y 24 no poseen muestras de todas las clases. Para este trabajo se consideraron los registros de los sujetos 1 a 20, 23, 25 y 26.

Haciendo uso de las herramientas `ffmpeg` y `ffprobe` desde `Python` se extrajeron los *frames* del tercio central de cada video componente. Debido a la gran similitud entre muestras contiguas, para este trabajo se conservó una muestra cada 5 *frames*. Una vez extraídos los *frames*, fue preciso asignar a cada imagen una etiqueta válida de salida de Deep Hand. En este sentido, se encontró que existían muestras con más de una etiqueta posible o con etiquetas inexistentes. A estas muestras se las asignó a la clase 0 –o *basura*– y se las descartó –clase 14–. En la Tabla C.2 del Anexo C se muestra la equivalencia de clases utilizadas para este trabajo. Las muestras de clases distintas de RWTH German Fingerspelling Database que poseían la misma etiqueta de salida de Deep Hand fueron conservadas. En adelante, se denominará 'DGS-FS' a la base de datos de imágenes conformada de la manera descripta en este apartado.

4.2.4. ASL Finger Spelling Dataset (ASL-FS)

ASL FingerSpelling Dataset es la base de datos de imágenes descripta en la Sec. 3.1.1. Debido a la gran similitud entre muestras contiguas, para este trabajo se consideró un subconjunto de Dataset A, conservando una muestra cada 10 imágenes. En este caso no fue necesario segmentar las imágenes. Para el etiquetado de los datos se empleó la Tabla C.3 del Anexo C. Las muestras asignadas a la clase 0 –o *basura*– fueron descartadas –clases 13 y 19–. Las muestras de clases distintas de ASL Finger Spelling Dataset que poseían la misma etiqueta de salida de Deep Hand fueron conservadas. En adelante, se denominará 'ASL-FS' a la base de datos de imágenes conformada de la manera descripta en este apartado.

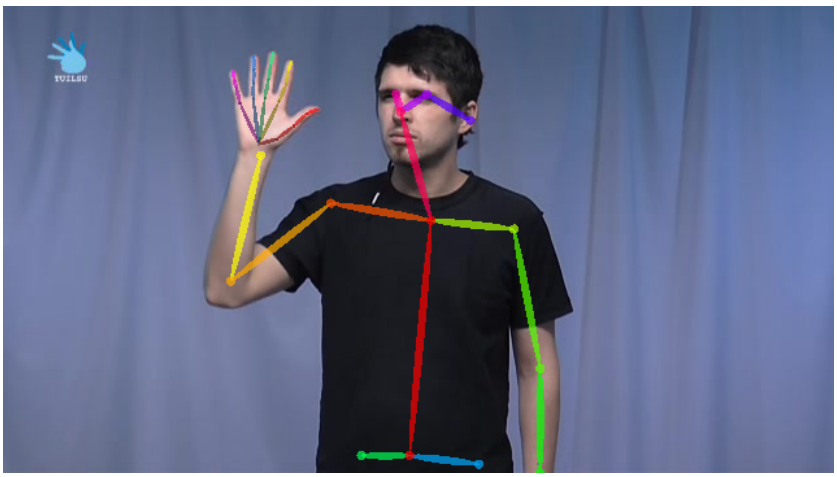
4.3. Preprocesamiento

Tal como se expuso en el apartado "Preprocesamiento de los datos" de la Sec. 4.1.2, Deep Hand requiere que la mano ocupe entre un tercio y la mitad del ancho/alto de la imagen de entrada. Luego, salvo para 'DH_Test', se hizo necesario realizar un *cropping* de las muestras a los fines de extraer un *patch* rectangular alrededor de

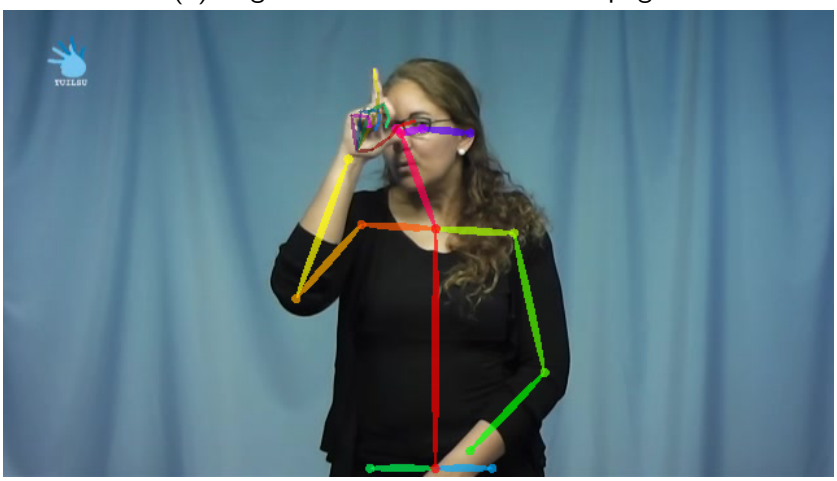
[9]Los registros provistos por la 'cámara 1' se descartaron por no cumplir los requerimientos de entrada a Deep Hand y por no ser fácilmente tratables con los métodos de preprocesamiento escogidos –ver la Sec. 4.3–.

una de las manos según los requerimientos mencionados. Para llevar a cabo este procesamiento se optó por el uso de OpenPose, fundamentalmente por su robustez frente a distintos tipos de fondo y por no requerir el uso de guantes por parte de los señantes.

Tal como se introdujo en la Sec. 2.6, OpenPose es una librería de acceso libre para la extracción de *keypoints* representativos de la postura corporal, de la actividad del rostro y de las manos de uno o más sujetos en una imagen RGB aislada. En la Figura 4.7 se muestran las estimaciones de las posturas corporal y manual provistas por OpenPose sobre dos muestras de la base de datos TReLSU-HS.



(a) Registro '1_6_from-file-065_fr34.png'.



(b) Registro '2_2_from-file-037_fr21.png'.

Figura 4.7: Registros de TReLSU-HS procesados mediante OpenPose.

Para el correcto uso de OpenPose fue necesario comprender con mayor detalle el formato de salida de la información deseada. A continuación se describen brevemente los modelos empleados internamente por OpenPose y las estrategias de *cropping* propuestas durante este trabajo.

Estimación de la postura corporal. Para la estimación de la postura corporal OpenPose ajusta un modelo interno de N nodos. La posición de cada uno de estos nodos se estima mediante un mapa denso de confianza provisto por una CNN específica [139].

En la Figura 4.8 se presenta el modelo de la postura corporal utilizado durante este trabajo, donde puede apreciarse el identificador correspondiente a cada uno de los 25 nodos detectados.

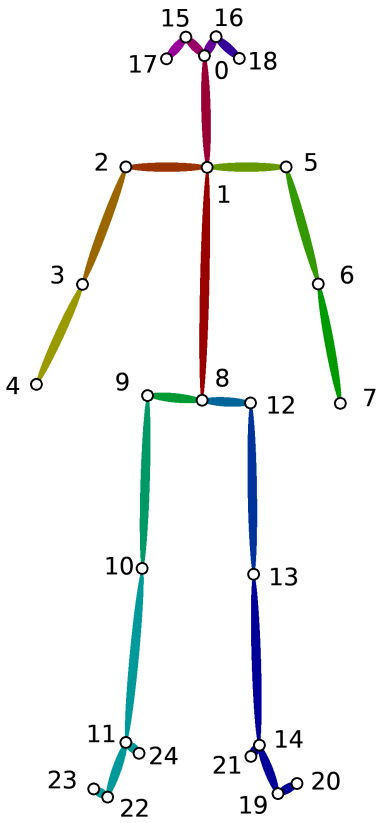


Figura 4.8: Modelo interno de 25 nodos empleado por OpenPose para estimar la postura corporal. Modificada de [1].

Estimación de la postura manual en 3D. Mediante el uso de un módulo dedicado –en adelante llamado módulo 'hand'–, OpenPose es capaz de estimar la posición 3D de las articulaciones de la mano, incluso en casos de auto-oclusión, tal como puede observarse en la mano de la Figura 4.7b. Para ello, dicho módulo emplea internamente una CNN entrenada con imágenes de múltiples perspectivas de la mano empleadas de forma conjunta mediante operaciones de triangulación y reproyección [119].

En la Figura 4.9 se muestra el modelo de 21 nodos ajustado sobre cada mano, donde pueden apreciarse la disposición e identificación de los *keypoints* detectados.

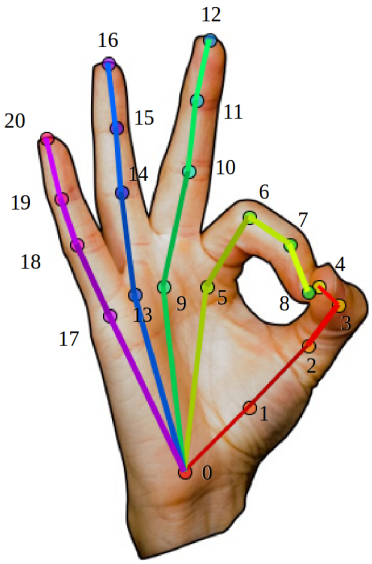


Figura 4.9: Modelo interno de 21 nodos empleado por el módulo 'hand' de OpenPose para estimar la postura de cada mano. Tomada de [1].

En base a las detecciones realizadas por OpenPose, se proponen *dos* estrategias de *cropping*:

- *Cropping* de la mano a partir de los *keypoints* del codo y muñeca derechas –*keypoints* 3 y 4 de la Figura 4.8–, en adelante llamado *cropping* 1.
- *Cropping* de la mano a partir de los *keypoints* detectados sobre la mano –Figura 4.9–, en adelante llamado *cropping* 2.

Cropping 1. Siguiendo una idea similar a 'Hand Bounding Box Detection' en [119], el *centro* de la mano en el plano imagen se estimó como:

$$\tilde{\mathbf{x}}_{h,1} = \mathbf{x}_b^{(4)} + \alpha \left(\mathbf{x}_b^{(4)} - \mathbf{x}_b^{(3)} \right),$$

donde $\mathbf{x}_b^{(3)}$ y $\mathbf{x}_b^{(4)}$ son las coordenadas 2D estimadas por OpenPose para los *keypoints* correspondientes al codo y a la muñeca *derechas*, respectivamente.

En cuanto a la extensión del *bounding box*, siguiendo una idea similar a 'Hand Bounding Box Detection' en [119], se propuso el siguiente ancho de *patch* $W_{\mathrm{P},1}$:

$$W_{\mathrm{P},1} = \left\lfloor \beta \left\| \mathbf{x}_b^{(1)} - \mathbf{x}_b^{(0)} \right\|_2 \right\rceil,$$

con $\lfloor \cdot \rceil$ denotando la función 'entero más cercano', y donde β es un margen de seguridad y $\mathbf{x}_b^{(1)}$ y $\mathbf{x}_b^{(0)}$ son los *keypoints* correspondientes a la base del cuello y a la nariz del señante, respectivamente. Considerando los requerimientos de Deep

Hand, se tomó $\beta \sim \mathcal{U}(a,b)$, siendo $\mathcal{U}(a,b)$ una variable aleatoria con distribución uniforme entre a y b. Los valores de a y b empleados se especifican para cada base de datos al final de la presente sección.

La altura $H_{\mathrm{P},1}$ del *patch* se determinó a partir de $W_{\mathrm{P},1}$ y de la relación de aspecto de las imágenes de la base de datos de prueba de Deep Hand –Sec. 4.2.1–, esto es:

$$H_{\mathrm{P},1} = \left\lfloor \frac{33}{23} W_{\mathrm{P},1} \right\rceil .$$

Cropping 2. En la estrategia basada en los *keypoints* de la mano, el *centro* de la mano en el plano imagen se estimó como:

$$\tilde{\mathbf{x}}_{h,2} = \frac{1}{6} \cdot \left[\mathbf{x}_h^{(0)} + \mathbf{x}_h^{(1)} + \mathbf{x}_h^{(5)} + \mathbf{x}_h^{(9)} + \mathbf{x}_h^{(13)} + \mathbf{x}_h^{(17)} \right],$$

donde $\left\{ \mathbf{x}_h^{(i)},\ \text{para } i = 0, 1, 5, 9, 13, 17 \right\}$ son las coordenadas 2D de los *keypoints* detectados sobre la *palma* de la mano por el módulo 'hand' de OpenPose.

En cuanto a la extensión del *bounding box*, se propuso el siguiente ancho de *patch* $W_{\mathrm{P},2}$:

$$W_{\mathrm{P},2} = \left\lfloor \beta \max_{i \neq j} \left\| \mathbf{x}_h^{(i)} - \mathbf{x}_h^{(j)} \right\|_2 \right\rceil ,$$

con $\beta \sim \mathcal{U}(a,b)$. Nuevamente, $H_{\mathrm{P},2} = \left\lfloor \frac{33}{23} W_{\mathrm{P},2} \right\rceil$. Esta solución se propuso para el caso en que el eje mayor de la mano poseyera una proyección mayor sobre la dirección horizontal.

Para el caso en que el eje mayor de la mano poseyera una proyección mayor sobre la dirección vertical se realizó:

$$H_{\mathrm{P},2} = \left\lfloor \beta \max_{i \neq j} \left\| \mathbf{x}_h^{(i)} - \mathbf{x}_h^{(j)} \right\|_2 \right\rceil ,$$

con $\beta \sim \mathcal{U}(a,b)$, resultando en este caso $W_{\mathrm{P},2} = \left\lfloor \frac{23}{33} H_{\mathrm{P},2} \right\rceil$. Los valores de a y b empleados se especifican para cada base de datos al final de la presente sección.

Tanto para *cropping 1* como para *cropping 2*, se realizó *zero-padding* para completar el *patch* sólo en aquellas muestras en que la mano se hallaba muy próxima a la frontera de la imagen. Ambos métodos de *cropping* probaron ser buenos y relativamente robustos sobre las bases de datos empleadas. Lógicamente, el método *cropping 2* posee un costo computacional mayor, por requerir de la estimación *a priori* de la postura corporal.

Para concluir el preprocesamiento, a cada imagen segmentada se le removió la *media por píxel*, según los requerimientos de Deep Hand expuestos en el apartado "Preprocesamiento de los datos" de la Sec. 4.1.2.

Las 2 estrategias de preprocesamiento expuestas en la presente sección se emplearon para el *cropping* de las muestras de TReLSU-HS y de DGS-FS. A continuación se identifican las bases de datos empleadas para la evaluación de Deep Hand en los experimentos expuestos en el Capítulo 5:

- DH_Test: base de datos de prueba de Deep Hand.
- TReLSU-HS_1: base de datos TReLSU-HS segmentada mediante *cropping 1*, con $\alpha = 0.15$ y $\beta \sim \mathcal{U}(1.75, 2.25)$.
- TReLSU-HS_2: base de datos TReLSU-HS segmentada mediante *cropping 2*, con $\beta \sim \mathcal{U}(2.25, 2.75)$.
- DGS-FS_1: base de datos DGS-FS segmentada mediante *cropping 1*, con $\alpha = 0.15$ y $\beta \sim \mathcal{U}(2.25, 2.75)$.
- DGS-FS_2: base de datos DGS-FS segmentada mediante *cropping 2*, con $\beta \sim \mathcal{U}(2.75, 3.25)$.
- ASL-FS: base de datos cruda.

4.4. Variantes de características para el aprendizaje por transferencia

Durante la resolución de un problema particular suele ser difícil contar con una base de datos lo suficientemente grande para entrenar una CNN "desde cero", esto es, ajustar una red inicializada con pesos aleatorios o nulos. Una estrategia muy empleada para ello es el *aprendizaje por transferencia* –o *transfer learning*–. Tal como su nombre lo indica, esta estrategia consiste en la transferencia del aprendizaje desde un "sistema base" hacia un "sistema objetivo". En primer lugar, se entrena un sistema para llevar a cabo una tarea "base" sobre un conjunto de datos "base". Luego, el sistema objetivo "toma" el conocimiento aprendido por el sistema base para llevar a cabo una tarea similar sobre un conjunto de datos "objetivo", generalmente mucho más pequeño [144]. En particular, los esquemas para implementar esta estrategia pueden ser [7]:

- hacer uso de modelos pre-entrenados de CNNs de aplicación general;
- realizar ligeras modificaciones sobre la arquitectura en las últimas capas de una CNN pre-entrenada y ajustar selectivamente los pesos de las nuevas capas a partir de una base de datos representativa del problema a resolver; o,
- emplear una CNN pre-entrenada como una etapa de extracción de características.

Durante este trabajo de maestría se hizo uso de la última de estas estrategias. En particular, a partir de las características extraídas por una CNN se llevó a cabo el entrenamiento de un clasificador SVM sobre las bases de datos DGS-FS y TReLSU-HS. Los detalles de estas pruebas y sus resultados se presentarán en la Sec. 5.4. A continuación se explican las dos variantes de extracción de características probadas.

4.4.1. Activaciones de la última capa oculta como características

Mediante esta estrategia se hizo uso de las redes Deep Hand e Inception-v3 como extractores de características. En particular, las características extraídas fueron los vectores de entrada a la última capa totalmente conectada de cada una de estas redes. En esta variante se extrajo de cada imagen un vector de N dimensiones que contiene las activaciones –salida de las ReLUs correspondientes– de las neuronas de la última capa oculta previa al clasificador [7]. Debido a la arquitectura de las redes, se tiene $N = 1024$ para Deep Hand y $N = 2048$ para Inception-v3. Se destaca que Inception-v3 fue entrenada a partir de la base de datos ImageNet, una base de datos compuesta para el reconocimiento de objetos de categorías muy diversas [126].

Para llevar a cabo la extracción de dichos vectores, se hizo uso de la función `register_forward_hook` de `PyTorch` y de la red Inception-v3 pre-entrenada disponible en `torchvision`. De forma previa a su ingreso a la red, cada muestra de DGS-FS y de TReLSU-HS fue preprocesada según se explicó en la Sec. 4.3.

4.4.2. *Keypoints* de la mano como características

En esta estrategia se propone hacer uso de una CNN entrenada para la detección de la posición 2D de las articulaciones manuales y digitales –*keypoints* de la mano–. Esto es, se propone el empleo de un vector de características compuesto por las coordenadas 2D de los *keypoints* detectados sobre la mano.

En particular, para realizar esta tarea se hizo uso del módulo 'hand' de OpenPose, el cual provee una descripción de la mano en función de 21 *keypoints* 2D –ver Figura 4.9–. A los fines de hacer un correcto uso de estas características, para cada *frame* se llevó a cabo la normalización de los *keypoints* detectados, de modo que el eje mayor de la mano quede dirigido verticalmente con la punta de los dedos hacia arriba. El eje mayor de la mano se estimó uniendo el *keypoint* 0 con el promedio de los *keypoints* 5, 9, 13 y 17 –ver Figura 4.9–. Finalmente, se normalizó el rango espacial de los *keypoints* para que se encontrara en el rango $[-1; 1]$, tanto en el sentido vertical como horizontal. En particular, OpenPose se corrió sobre los *frames* crudos de las bases de datos DGS-FS y TReLSU-HS.

Durante este trabajo de maestría, el módulo 'hand' se empleó de forma dependiente de la detección de la postura corporal. Luego, no fue posible procesar la base de datos ASL-FS por no mostrar visible el cuerpo de los señantes en sus muestras. Si bien existe una implementación *stand alone* del módulo 'hand' no fue posible su reproducción en el marco de este trabajo[10].

Por último, se mencionan dos herramientas invocadas desde OpenCV, las cuales han quedado pendientes de explorar. Por un lado, la implementación *stand alone* del módulo 'hand' de OpenPose invocado desde OpenCV[11]. De esta manera

[10] https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/standalone_face_or_hand_keypoint_detector.md.

[11] https://www.learnopencv.com/hand-keypoint-detection-using-deep-learning-and-opencv/

será posible la extracción de *keypoints* de la mano sobre bases como ASL-FS y, eventualmente, la exploración de soluciones completamente independientes de la etapa de *cropping* o segmentación. Por otro lado, una librería alternativa a OpenPose denominada `wrnchAI`[12], la cual se invoca desde OpenCV y posee las mismas funcionalidades pero mejores capacidades de resolver ambigüedades en algunos casos de oclusión y una velocidad de cómputo de 2 a 4 veces mayor, dependiendo del tamaño de la imagen de entrada.

4.5. Comentarios de fin de capítulo

A lo largo de este capítulo se describieron los elementos necesarios para llevar a cabo los experimentos presentados en el Capítulo 5. En primer lugar, se explicaron los fundamentos del sistema Deep Hand, los detalles de implementación por parte de sus autores y los detalles de su reproducción en el marco de este trabajo. Luego, se expusieron las bases de datos empleadas y su conformación. En tercer lugar, se presentaron dos estrategias basadas en aprendizaje profundo para llevar a cabo el preprocesamiento requerido por Deep Hand. Por último, se presentó una manera de emplear OpenPose y Deep Hand como extractores de características a los fines de realizar aprendizaje por transferencia.

[12]https://www.learnopencv.com/pose-detection-comparison-wrnchai-vs-openpose/

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 5

Experimentos y resultados

En este capítulo se describen los experimentos realizados durante este trabajo de tesis. En la Sec. 5.1 se presenta una evaluación de las salidas de Deep Hand frente a las distintas bases de datos consideradas. En la Sec. 5.2 se evalúa la respuesta de Deep Hand frente a la base de datos ASL-FS a distintos niveles de *zero-padding*. Luego, en la Sec. 5.3 se muestra una comparación de las características provistas por Deep Hand y por Inception-v3 sobre las bases de datos consideradas. Por último, en la Sec. 5.4 se evalúa el desempeño de un clasificador SVM sobre las bases de datos DGS-FS y TReLSU-HS a partir de las características extraídas por Deep Hand y OpenPose bajo un esquema de aprendizaje por transferencia.

5.1. Evaluación de las salidas de Deep Hand

Mediante este experimento se evaluaron las salidas de Deep Hand. Para una interpretación más ágil, en las figuras de esta sección se muestran las matrices de confusión *normalizadas* y las *accuracies* top-1 y top-5 correspondientes a cada una de las bases de datos tratadas. La normalización de las matrices de confusión se llevó a cabo por fila, esto es, para cada clase deseada $\hat{c}$ la cantidad de salidas correctas e incorrectas se normalizó frente la cantidad de muestras pertenecientes a la clase $\hat{c}$. Luego, la lectura de la escala de colores de la representación de las matrices debe realizarse por fila. En este sentido, debe tenerse cuidado con la lectura de la escala de colores de la matriz normalizada entre las distintas columnas, puesto que no siempre se cuenta con una base de datos balanceada, esto es, con la misma cantidad de muestras de cada clase. En el Anexo B se reportan las matrices de confusión crudas, esto es, cada entrada de la matriz está expresada en cantidad de muestras.

En primer lugar, se evaluó el desempeño del sistema implementado en `PyTorch` sobre la base de datos de prueba DH_Test. Asimismo, se comparó el desempeño de este sistema frente a la implementación en `TensorFlow` de Camgoz introducida en la Sec. 4.1.3. En la Figura 5.1 se observan las matrices de confusión correspondientes. En términos generales, se observa un muy buen desempeño, 84.88 % *versus* 85.44 % de *accuracy* top-1, con algunos elementos fuera de la diagonal.

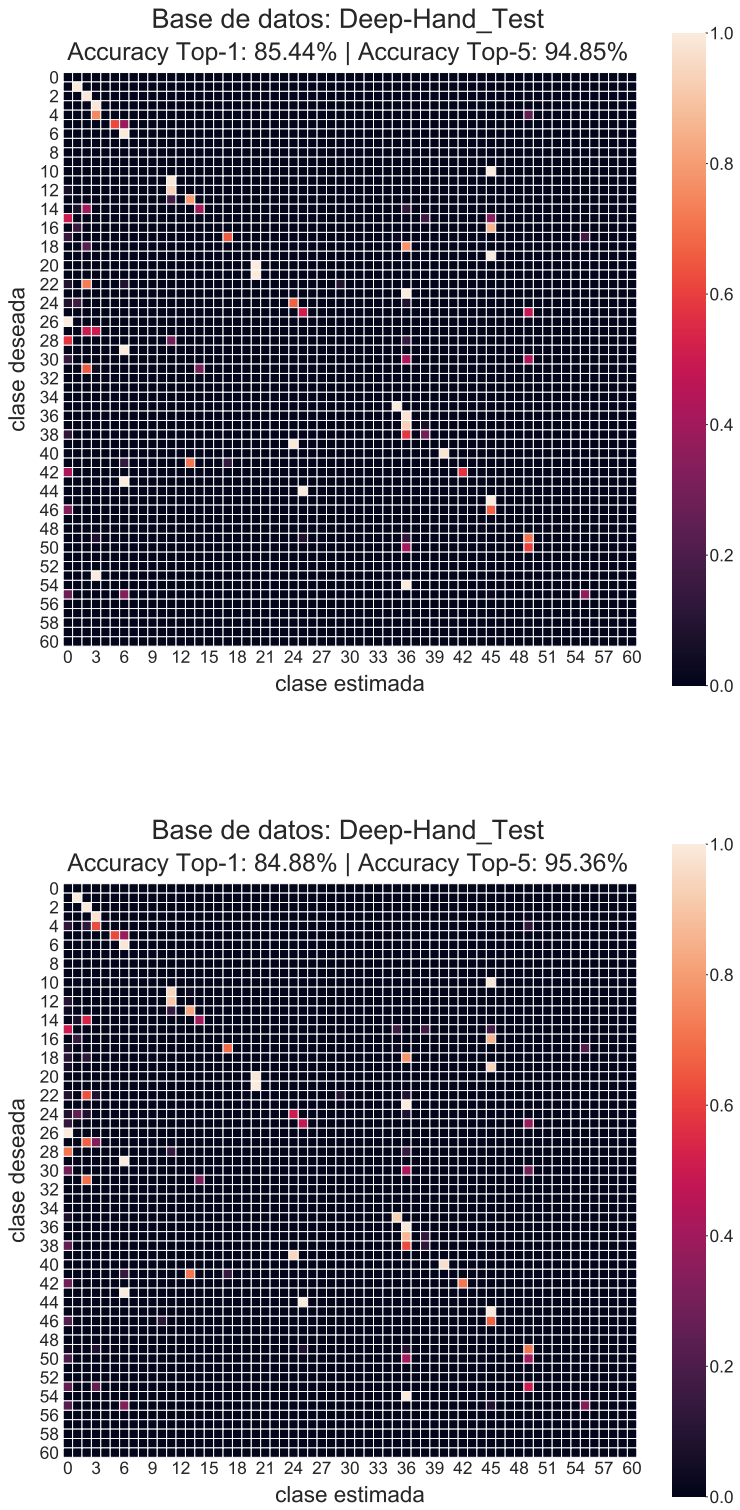


Figura 5.1: Matrices de confusión normalizadas del sistema Deep Hand frente a la base de datos DH_Test. Implementación de Camgoz (arriba) e implementación en `PyTorch` (debajo).

En este punto cabe aclarar que la base de datos DH_Test es una base de datos no balanceada, esto es, no posee la misma cantidad de muestras por clase. Por esta razón debe tenerse cuidado con la lectura de la escala de colores, tal como se comentó. No obstante, a partir de la Figura 5.1, es posible observar que prácticamente todas las muestras de las clases 10, 19, 21, 23, 26, 29, 43, 44 y 54 fueron respectivamente confundidas con las clases 45, 45, 20, 36, 0, 6, 6, 25 y 36. Tomando en consideración la Tabla A.1 del Apéndice A, se puede observar que el sistema falla ante entradas muy similares,  (10) o  (19) *versus*  (45),  (21) *versus*  (20)–, de forma parcialmente consistente con lo reportado por Koller y cols. en [75].

A partir de la Figura 5.1 es posible observar que las clases *mejor* clasificadas son las 13 clases reportadas en la Tabla 3 de [75]. No fue posible reproducir los valores de *accuracy* global top-1 de 62.8 % y top-5 de 85.6 % a partir de las matrices reportadas en [75]. Aparentemente, este desempeño correspondía al primer modelo implementado el cual ya no se encuentra disponible –`1-miohands-v1` reportado por Koller y cols. en [4]–.

En cuanto a la implementación de Camgoz, se observa que el valor de 85.44 % de *accuracy* top-1 fue obtenido como la proporción de clasificaciones correctas sobre el total de entradas. Asimismo, a partir de los *puntajes* –o *scores*– obtenidos mediante el *script* `evaluation.py` se calculó el valor de *accuracy* top-5, resultando en un 94.85 %. Ambos valores son consistentes con el desempeño del modelo `1-miohands-v2` reportado por Koller y cols. en [4].

En este punto es conveniente aclarar que las medidas de *accuracy* así obtenidas constituyen estimadores sin sesgo de la probabilidad de clasificar correctamente una nueva muestra, siempre y cuando la base de datos empleada se encuentre balanceada. En la Figura 5.2 se muestra la distribución de muestras por clase de las 45 clases de la base de datos DH_Test. Las clases 0, 7, 8, 9, 32, 33, 34, 47, 48, 51, 52, 56, 57, 58, 59 y 60 no se encuentran muestreadas, y por tanto no se han incluido en la figura.

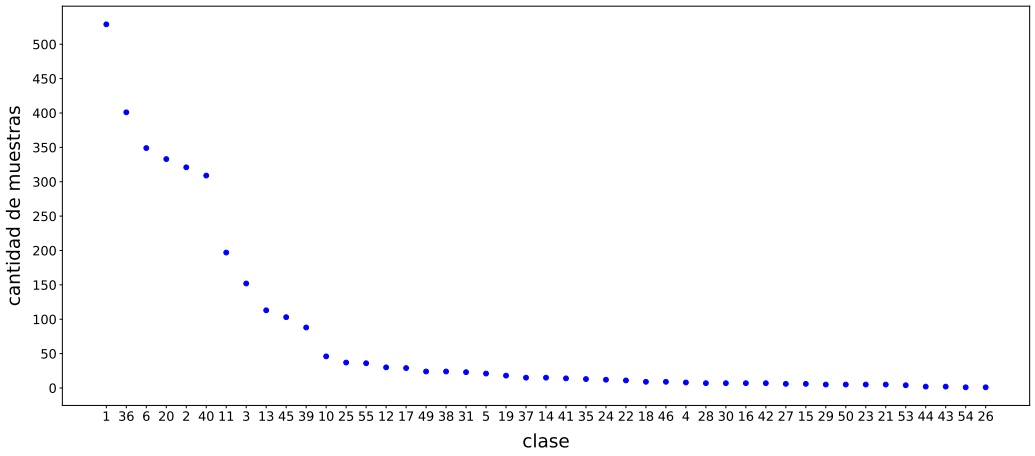


Figura 5.2: Distribución de muestras por clase para la base de datos DH_Test.

En base a lo observado, se tiene una base de datos con un gran desbalance, incluso con clases no muestreadas. Por lo tanto, las medidas de *accuracies* reportadas por Camgoz y Koller no son representativas del desempeño del sistema frente a las 45 clases consideradas. En adelante se referirá como $accuracy_{\mathrm{B}}$ a la medida de *accuracy* top-1 que supone una base de datos balanceada. Asimismo, se propone la medida $accuracy_{\mathrm{UB}}$ para estimar la *accuracy* top-1 de un sistema de clasificación de N clases sobre una base de datos desbalanceada, a calcular como:

$$accuracy_{\mathrm{UB}} = \frac{\mathrm{Tr}\{\mathrm{MCN}\}}{N_{\mathrm{SC}}},$$

siendo N_{SC} la cantidad de clases efectivamente muestreadas en la base de datos empleada y MCN la matriz de confusión normalizada por 'clase deseada', según se explicó al comienzo de esta sección.

De esta manera, se tiene un $accuracy_{\mathrm{UB}}$ de 34.39 % y de 33.64 % sobre DH_Test para la implementación de Camgoz y en `PyTorch`, respectivamente. Los resultados reportados en adelante fueron obtenidos a partir de la implementación en `PyTorch`.

Para el resto de las bases de datos el sistema muestra desempeños muy pobres, valores de *accuracies* inferiores al 20 %. En las Figura 5.3, 5.4 y 5.5 se presentan las matrices de confusión normalizadas del sistema Deep Hand frente a las bases de datos ASL-FS, TReLSU-HS y DGS-FS, respectivamente.

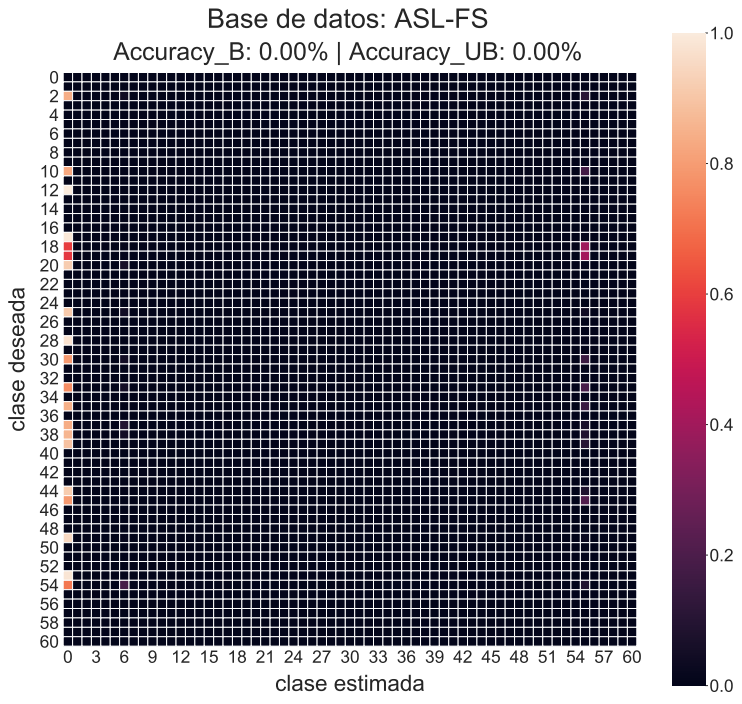


Figura 5.3: Matrices de confusión normalizadas del sistema Deep Hand.

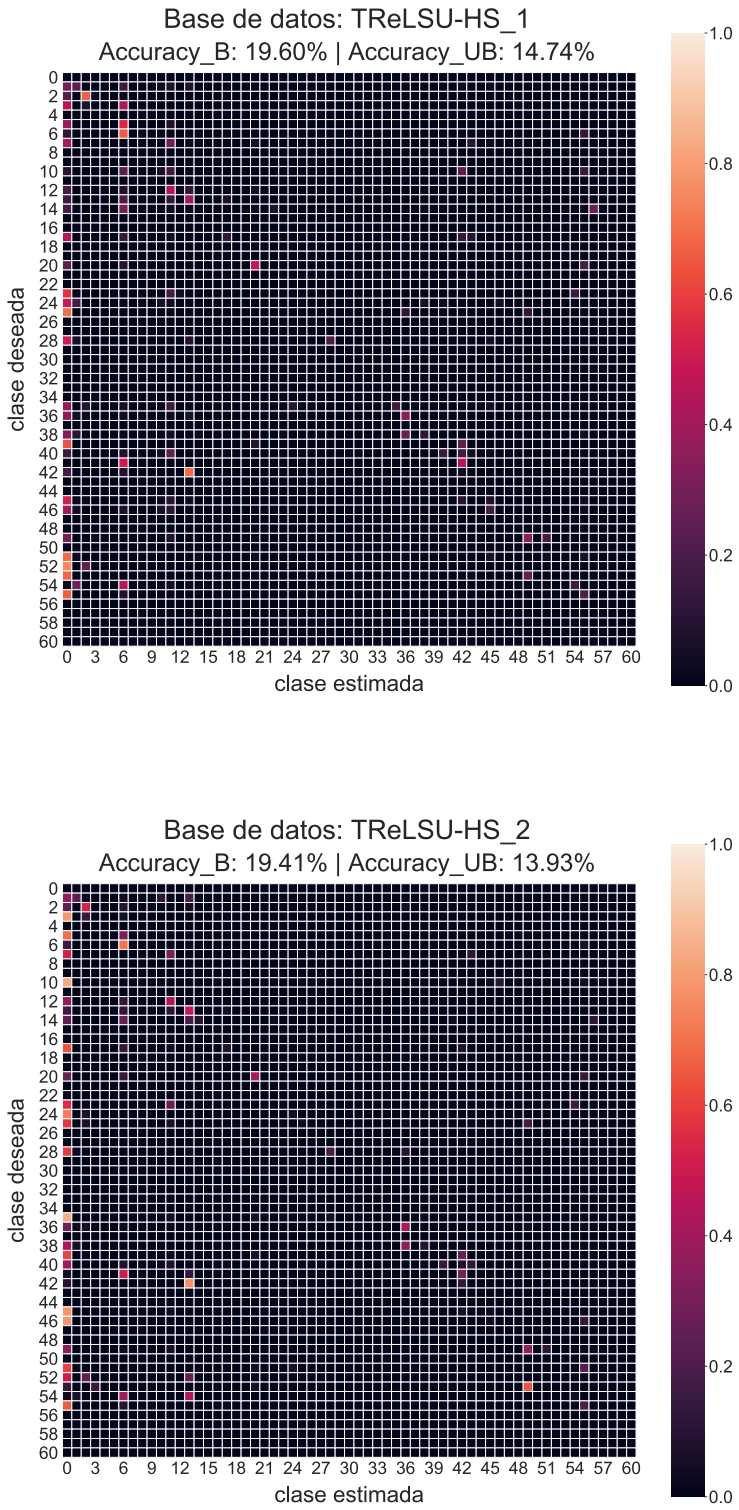


Figura 5.4: Matrices de confusión normalizadas del sistema Deep Hand.

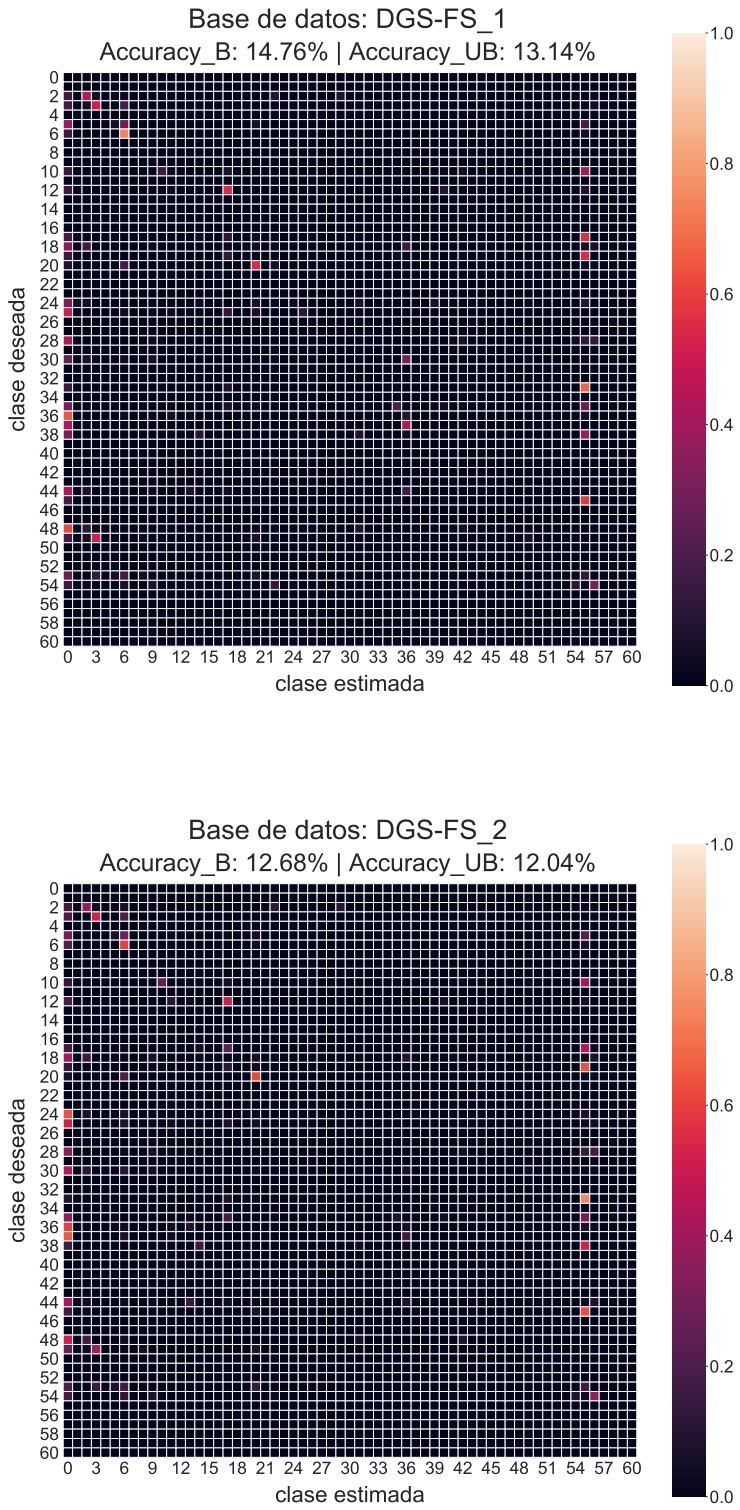


Figura 5.5: Matrices de confusión normalizadas del sistema Deep Hand.

Analizando las matrices de confusión de las Figura 5.3, 5.4 y 5.5, es posible afirmar que el sistema reconoce relativamente bien sólo unas pocas clases –2, 3, 6 y 20– y posee una gran parte de las muestras asignadas a las clases 0 –o *basura*– y 55 –sólo sobre DGS-FS–.

El bajo desempeño del sistema Deep Hand sobre bases de datos "ajenas" podría deberse a varias razones.

Por un lado, a una segmentación incorrecta de las imágenes. Considerando la matriz de confusión de la base de datos ASL-FS –Figura 5.3–, es posible observar que la gran mayoría de las muestras fueron clasificadas como clase 0 –o *basura*–, esto es, el sistema no es capaz de diferenciar las muestras. Este hecho podría deberse a que la mano no posee el tamaño requerido por Deep Hand –en ASL-FS la mano ocupa prácticamente toda la imagen de entrada–. En el experimento de la Sec. 5.2 se explora la dependencia del *accuracy* frente a bordes de *zero-padding* de distinto espesor sobre las imágenes de esta base de datos.

Por otro lado, tomando en consideración las Figura 5.4, 5.5 y 5.3, la asignación de etiquetas a las bases de datos de acuerdo a las Tablas C.1, C.2 y C.3 del Anexo C podría no haber sido óptima en cuanto a las salidas de Deep Hand. No obstante, considerando que la gran parte de las muestras mal clasificadas fueron asignadas a la clase 0 –o *basura*–, es posible afirmar que el etiquetado de las bases no es el problema mayor. Por completitud, en el Anexo D se muestran las matrices de confusión de Deep Hand sobre las bases de datos etiquetadas según las clases de origen de cada base de datos, esto es, ignorando las tablas de equivalencias del Anexo C.

Finalmente, la estimación de la media por píxel que se remueve a cada imagen de entrada podría estar sesgada. En el Anexo E se incluyen las matrices de confusión del sistema empleando las imágenes de entrada sin la remoción de la media. Sorpresivamente, salvo para la base de datos DH_Test, se observa una mejora apreciable en el desempeño en todos los casos. Este hecho sugiere que la estimación de la media es un aspecto importante en la clasificación. Por motivos de tiempo, ha quedado fuera de los alcances de este trabajo la exploración de distintas estrategias para mejorar la estimación de la media de cada base de datos.

5.2. *Zero-padding* sobre ASL-FS

Mediante este experimento se buscó evaluar el desempeño del sistema frente a la mano segmentada a distintas escalas en la imagen. Para ello, de manera previa al remuestreo de las imágenes se efectuó un *zero-padding* –esto es, la *adición de ceros*– de distintos espesores a los lados de la imagen. El *zero-padding* se realizó de manera proporcional a d, siendo d la menor de las dimensiones de la imagen; agregando $\lfloor p \cdot d \rceil$ ceros a un lado y otro de la imagen en la dirección de d y agregando ceros en la dimensión restante de modo que la imagen resultante tenga una relación de aspecto de $\frac{23}{33}$, según se comentó en la Sec. 4.2.1. En la Figura 5.6 se observa una muestra de la base de datos ASL-FS, a la izquierda la muestra original y a la derecha con tres niveles distintos de *zero-padding*.

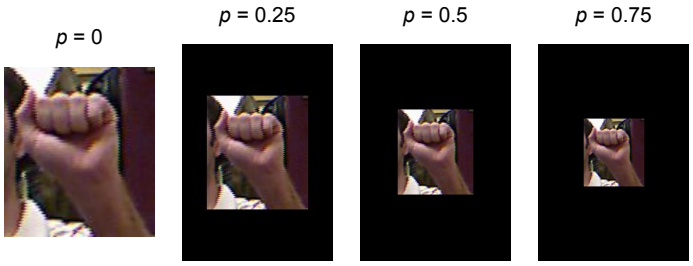


Figura 5.6: Distintos niveles de *zero-padding* sobre una muestra de la base de datos ASL-FS.

En la Figura 5.7 se observa el porcentaje de $accuracy_{\mathrm{UB}}$ conforme varía p. En términos generales se observa que el desempeño global es muy malo independientemente de p. En el Anexo F se adjuntan las matrices de confusión obtenidas durante este experimento. En la Figura 5.8 se observa la proporción de las salidas no clasificadas como clase 0 –o *basura*–, aquí denominada 'NZ_outputs'.

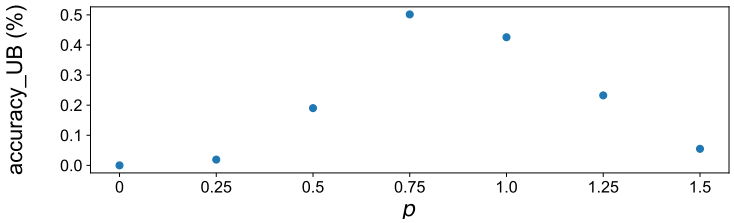


Figura 5.7: Desempeño del sistema Deep Hand sobre la base de datos ASL-FS a distintos niveles de *zero-padding*.

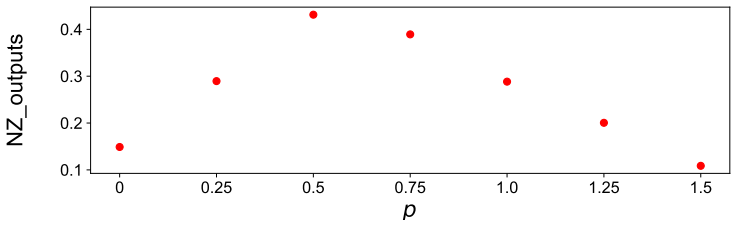


Figura 5.8: Proporción de salidas no asignadas a la clase 0 –o *basura*– del sistema Deep Hand sobre la base de datos ASL-FS a distintos niveles de *zero-padding*.

Analizando de manera conjunta las Figura 5.7 y 5.8 se observa que NZ_outputs presenta un máximo de 0.43, para $p = 0.5$, con un $accuracy_{\mathrm{UB}}$ correspondiente de 0.19 %. El $accuracy_{\mathrm{UB}}$ máximo se obtuvo para $p = 0.75$, con un valor de 0.5 %. Luego, se interpreta que el sistema no es capaz de clasificar bien las imágenes que efectivamente puede discriminar de la clase 0 –o *basura*–. Analizando las matrices de confusión del Anexo F para $p = [0.5, 0.75]$, se puede apreciar que el sistema confunde las clases detectadas –2, 18, 19, 20, 25, 28, 33, 37, 38, 44, 45 y 54– con sólo dos clases –6 y 55–. De esta manera se concluye además que la confusión observada no viene dada precisamente por la similitud interclase.

5.3. Deep Hand *versus* Inception-v3

Mediante este experimento se buscó la evaluación de dos aspectos de Deep Hand. Por un lado, qué tan próximas se encuentran las representaciones de cada una de las configuraciones manuales en el espacio de las características provistas por Deep Hand. Por otro lado, se buscó comparar la calidad de estas características frente a las características provistas por Inception-v3. Por su parte, Inception-v3 fue entrenada a partir de la base de datos ImageNet, una base de datos compuesta para el reconocimiento de objetos de categorías muy diversas [126]. Desde el punto de vista metodológico hubiera sido mejor comparar la calidad de Deep Hand contra Inception-v1, la red a partir de la cual se implementó Deep Hand. Sin embargo, no fue posible encontrar una implementación de Inception-v1 pre-entrenada para `PyTorch`.

Para la implementación de este experimento, se realizó la extracción de características de acuerdo a lo expuesto en la Sec. 4.4.1 y se implementó un clasificador por KNN, para $K = [1, 3, 7, 15, 30, 45, 60, 75]$.

El desempeño de los clasificadores se caracterizó mediante las dos técnicas complementarias descriptas a continuación:

- una validación cruzada de M *iteraciones*[1] –en adelante, VC–, técnica en la cual se toman M particiones aleatorias de los datos, empleando $\frac{2}{3}$ para el entrenamiento y el $\frac{1}{3}$ restante para el testeo de cada clasificador [73]. En particular, se tomó $M = 10$ para todas las bases de datos.

- una validación cruzada por sujeto –en adelante, VCS–, técnica en la cual se realiza el testeo del clasificador con los datos provenientes de un solo sujeto y el entrenamiento con los datos de los sujetos restantes. Este procedimiento se repitió para cada uno de los sujetos participantes de cada base de datos. Luego, se realizó 5 veces para TReLSU-HS, 5 para ASL-FS y 22 veces para DGS-FS. La base de datos Deep-Hand_Test –ver Sec. 4.2.1– no fue incluida en este procedimiento por no estar disponible el sujeto fuente de cada uno de los datos.

Vale aclarar que durante este experimento ASL-FS, DGS-FS_1 y DGS-FS_2 se emplearon con sus etiquetas de origen. En adelante se referirá como `'source_labels'` a este aspecto de cada base de datos.

En los diagramas de caja de la Figura 5.9 y en la Figura 5.10 se ilustra la distribución de $accuracy_{\mathrm{UB}}$ conforme varía K para los procedimientos de VC y VCS, respectivamente. En particular, para la VC –Figura 5.9– se observa que conforme aumenta K la *accuracy* decae, mostrando un comportamiento asintótico en algunos casos. El decaimiento se atribuye a que en cada iteración de la VC existen muestras muy similares en el conjunto de prueba, en particular aquellas provenientes del mismo sujeto en la misma sesión. Luego, una vez que las muestras similares se han agotado, la calidad de la clasificación comienza a decaer.

[1] Traducido del inglés, *fold*.

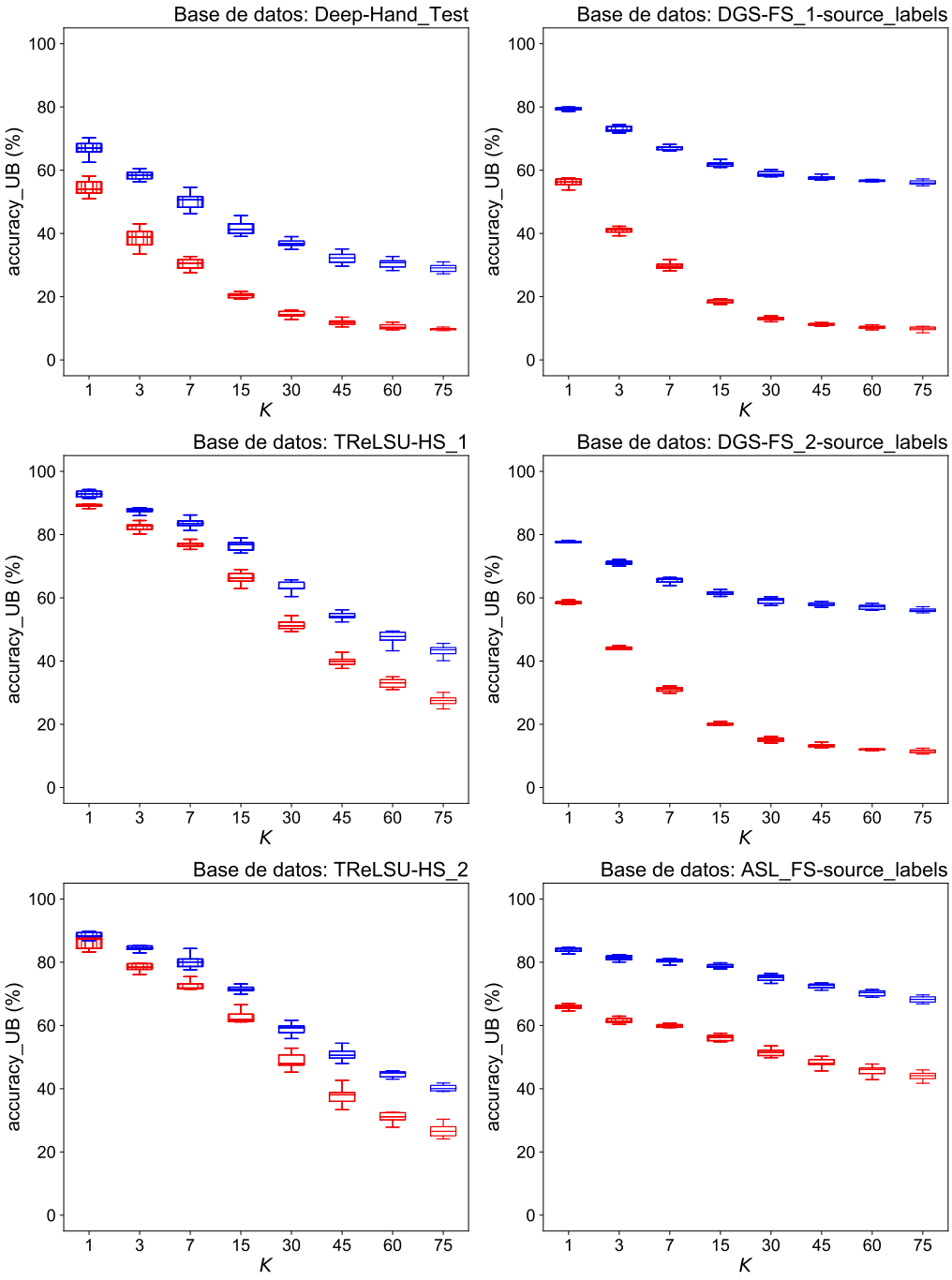


Figura 5.9: VC sobre las distintas bases de datos. Referencia de colores: Deep Hand en azul e Inception-v3 en rojo.

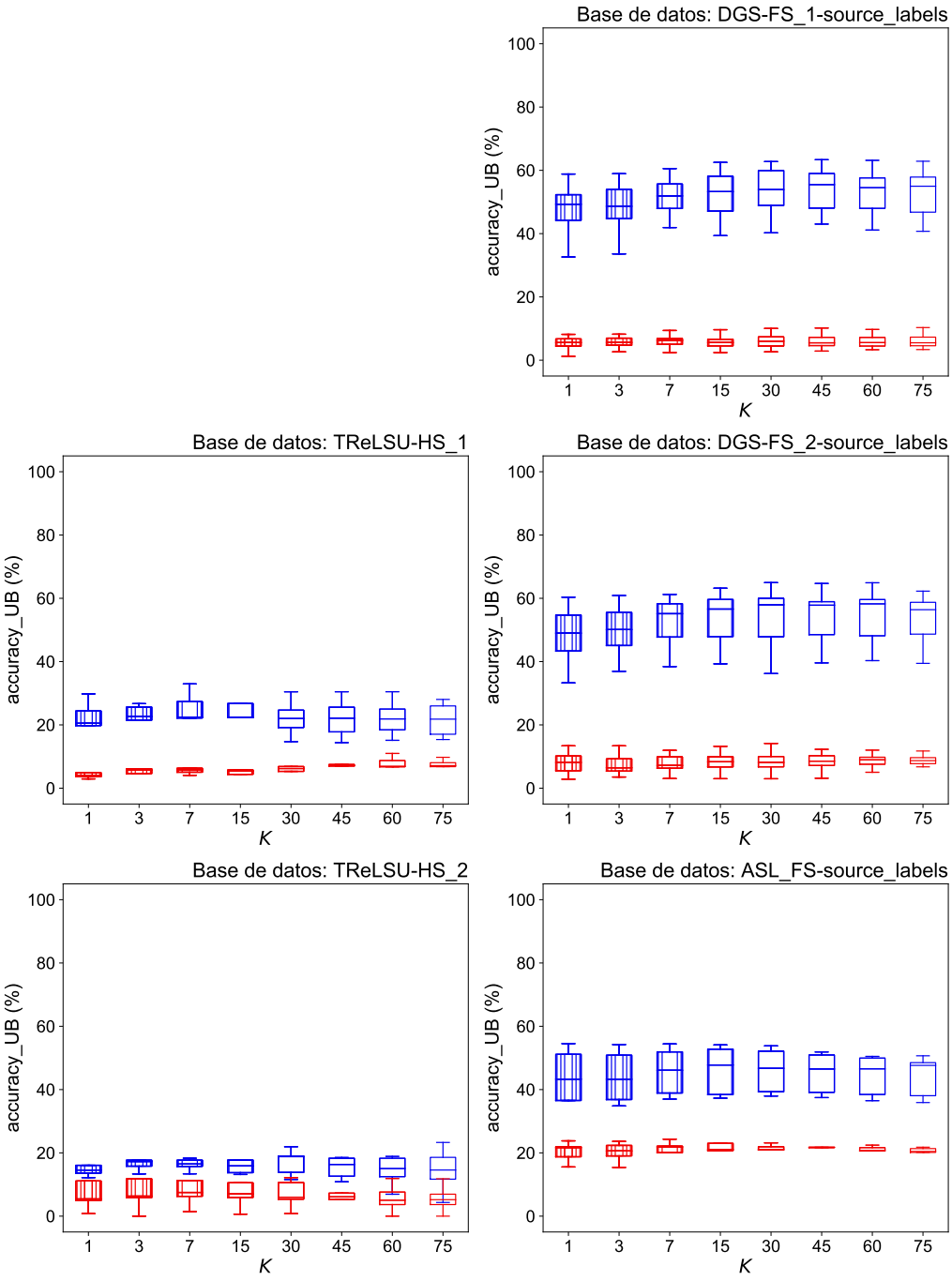


Figura 5.10: VCS sobre las distintas bases de datos. Referencia de colores: Deep Hand en azul e Inception-v3 en rojo.

Por otro lado, para la VCS –Figura 5.10– se observa un desempeño notablemente inferior de los clasificadores con respecto a VC. Tomando en consideración el diseño del experimento, la caída en el desempeño era esperable. Debido a la proveniencia de distintos sujetos los vectores de características empleados para el testeo podrían ser muy diferentes de aquéllos empleados durante el entrenamiento.

A la izquierda de la Figura 5.10 se observa un desempeño muy pobre sobre la base de datos TReLSU-HS. Este desempeño se atribuye al desbalance de TReLSU-HS –ver Figura 4.6–, habiendo iteraciones en las cuales no fue posible presentar durante el entrenamiento *todas* las clases a testear posteriormente. Si bien este error metodológico fue detectado, ha quedado fuera de los alcances de esta tesis el replanteo de la base de datos TReLSU-HS. Por este motivo, el uso de TReLSU-HS requirió de una estrategia particular en los experimentos subsecuentes.

En relación al primer aspecto planteado al comienzo de esta sección, es posible concluir que las características provistas por Deep Hand brindan una tasa de reconocimiento del orden del 50 % sobre las bases DGS-FS –35 clases– y ASL-FS –24 clases–. Este hecho sugiere que la representaciones de las distintas muestras de cada clase se encuentran relativamente próximas en el espacio de características. Ha quedado fuera de los alcances de este trabajo el análisis de este fenómeno mediante técnicas de *clustering*.

Por último, en relación al segundo aspecto planteado al comienzo de esta sección, a partir de este experimento es posible concluir que las características provistas por Deep Hand permiten una mejor discriminación de configuraciones manuales frente a las provistas por Inception-v3.

5.4. Aprendizaje por transferencia

En vista de las matrices de confusión y de las bajas *accuracies* para las bases de datos ajenas consideradas, se propuso el entrenamiento de un clasificador SVM a partir de las características según se comentó en la Sec. 4.4. En particular, durante este experimento de *aprendizaje por transferencia* se probaron cuatro variantes de características:

- 'cDH': vector de 1024 características provistas por Deep Hand de forma independiente, extraídas según se comentó en la Sec. 4.4.1.
- 'cOP': vector de 42 características, *keypoints* de la mano provistos por OpenPose de manera independiente, extraídos y normalizados según se comentó en la Sec. 4.4.2.
- 'cDHOP': concatenación de los vectores de características 'cDH' y 'cOP'.
- 'cInc-v3': vector de 2048 características provistas por Inception-v3 de forma independiente, extraídas según se comentó en la Sec. 4.4.1.

La última de estas variantes se incluyó a modo de referencia. Asimismo, en paralelo al entrenamiento del clasificador SVM, se llevó a cabo el entrenamiento de un clasificador por KNN, con $K = [1, 3, 7, 15, 30, 45, 60, 75]$.

A continuación se describe el uso de las bases de datos DGS-FS y TReLSU-HS durante este experimento en particular. Luego, en la Sec. 5.4.2 se comentan los detalles de implementación de los clasificadores utilizados.

5.4.1. Consideraciones sobre las bases de datos empleadas

Durante este experimento se trabajó bajo un esquema de VCS sobre las bases de datos DGS-FS_1, DGS-FS_2, TReLSU-HS_1 y TReLSU-HS_2. En particular, las dos variantes de DGS-FS se emplearon con sus etiquetas de origen –`'source_labels'`–, tal como se describió en la Sec. 4.2.3 y preprocesadas según se presentó en la Sec. 4.3. No obstante, sobre las dos variantes de TReLSU-HS fue necesario realizar las modificaciones que se describen en el párrafo siguiente.

En primer lugar, cabe recordar que TReLSU-HS no posee un muestreo uniforme de las clases sobre los distintos sujetos –ver Figura 5.11–. En este sentido, el esquema de VCS requirió descartar en primer lugar las muestras de sujeto único –marcadas en rojo en la Figura 5.11–. Luego, para cada clase y cada sujeto, se aislaron las muestras correspondientes como conjunto de testeo, empleando el resto como el conjunto de entrenamiento correspondiente. Lógicamente, en estas condiciones los conjuntos de entrenamiento y testeo poseen datos provenientes del mismo sujeto, por lo cual no se trata de un proceso de VCS en sentido estricto. Esta decisión se tomó para aprovechar mejor el reducido tamaño de TReLSU-HS –alrededor de 2800 muestras sobre 22 clases–.

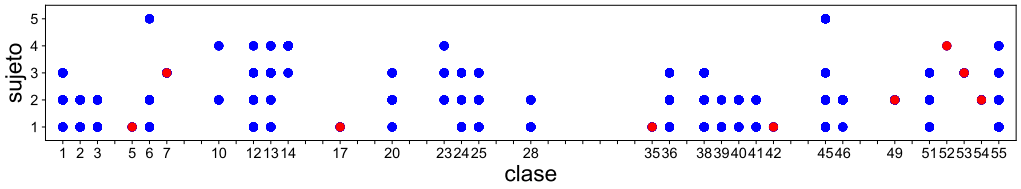


Figura 5.11: TReLSU-HS: Muestreo de las clases en los 5 sujetos etiquetados.

5.4.2. Implementación de los clasificadores

El fundamento de las dos técnicas de clasificación empleadas en este experimento fue expuesto en la Sec. 2.5.1. Para la implementación de los clasificadores por KNN y SVM se hizo uso de las funciones `neighbors.KNeighborsClassifier`[2] y `svm.SVC`[3], respectivamente, ambas del paquete `scikit-learn`.

En particular, `neighbors.KNeighborsClassifier` se usó con la distancia euclidea como métrica de distancia. Por su parte, `svm.SVC` se utilizó con $C = 1$ en la Ec.(2.9), empleando un *kernel* de tipo gaussiano –opción `'rbf'` de la implementación usada–, con $\gamma = N_F^{-1}$, siendo N_F la dimensión del vector de características;

[2]https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[3]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

esto es, 1024, 2048, 42 y 1066 para 'cDH', 'cInc-v3', 'cOP' y 'cDHOP', respectivamente. Ha quedado fuera del alcance de esta tesis una búsqueda exhaustiva de los parámetros C y γ del clasificador SVM. Vale agregar que la función `svm.SVC` realiza una clasificación multiclase bajo un esquema "uno *versus* uno". Siendo N_C la cantidad de clases a discriminar, bajo este esquema se realiza el entrenamiento de $N_C \cdot (N_C - 1)/2$ clasificadores, cada uno de los cuales discrimina el dato de entrada en dos clases.

5.4.3. Resultados sobre DGS-FS-source_labels

En la Figura 5.12 se presenta la medida de $accuracy_{\mathrm{UB}}$ del sistema frente a las bases DGS-FS_1 y DGS-FS_2, para las distintas variantes de características mencionadas. En línea sólida se presenta el desempeño del sistema basado en SVM y en línea de trazos el desempeño del sistema basado en KNN. En cuanto a este último, en la figura sólo se presenta el desempeño del sistema con máxima $accuracy_{\mathrm{UB}}$ media –ver en la leyenda el valor de K óptimo–.

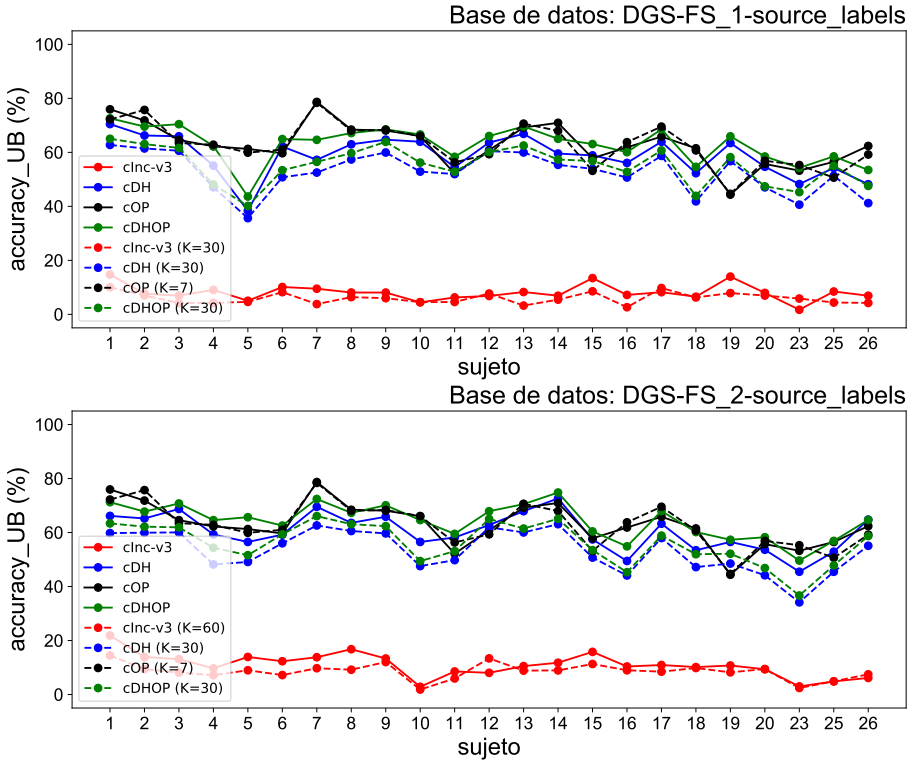


Figura 5.12: Entrenamiento de un clasificador: VCS sobre DGS-FS_1 y DGS-FS_2. Referencia: SVM en línea sólida y clasificador por KNN en línea de trazos.

Como puede observarse en la Figura 5.12 el desempeño global es notablemente superior al obtenido mediante el sistema Deep Hand "pre-entrenado", el cual mostró una $accuracy_{\mathrm{UB}}$ de 13.14 % y 12.04 % sobre DGS-FS_1 y DGS-FS_2, respectivamente –ver Figura 5.5–. De esta manera, bajo un esquema de aprendizaje por

transferencia se sacó provecho de Deep Hand y de OpenPose como extractores de características. No se observan diferencias apreciables entre los comportamientos 'cDH', 'cOP' y 'cDHOP'. En cuanto a la etapa de clasificación, se acusa un mejor desempeño del sistema basado en SVM frente al correspondiente por KNN sobre las características 'cDH' y 'cDHOP'. Cabe destacar que el vector de características 'cOP' es de 42 dimensiones, mientras que 'cDH' es de 1024. Además, para 'cOP' la clasificación por KNN es mejor o igual que SVM para la gran mayoría de los sujetos, independientemente del método de *cropping*.

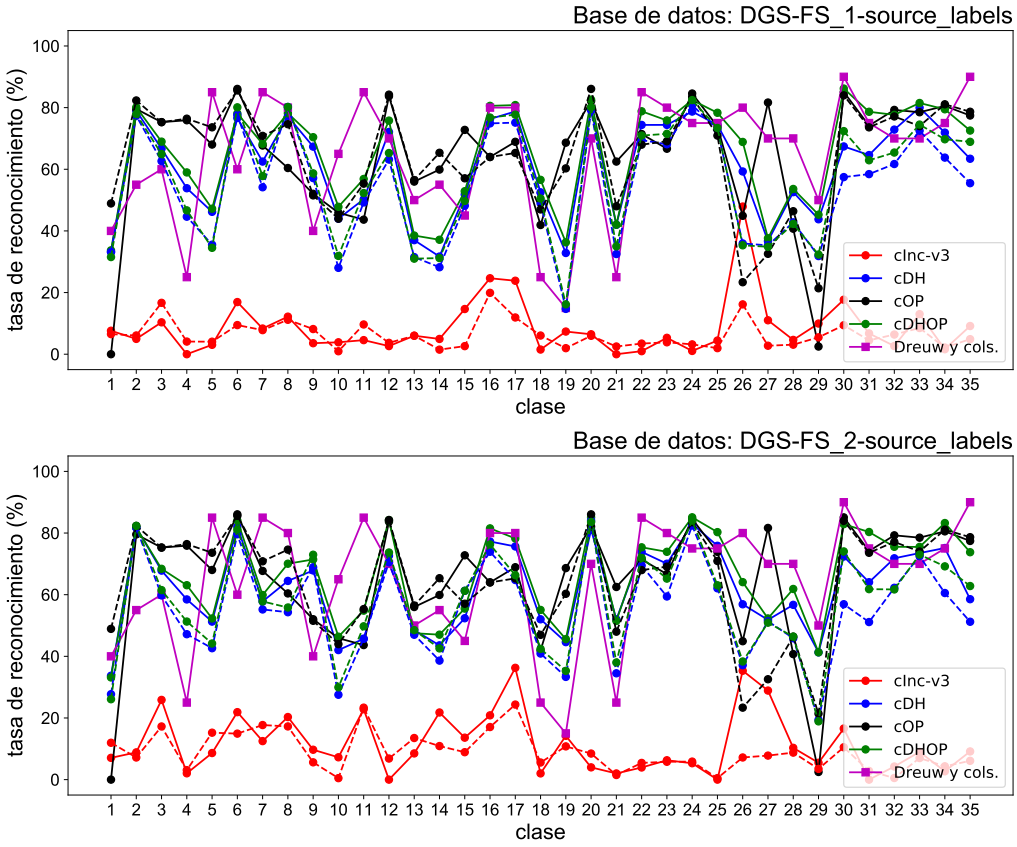


Figura 5.13: Tasa de reconocimiento por clase sobre DGS-FS. Referencia: SVM en línea sólida y clasificador por KNN en línea de trazos.

En la Figura 5.13 se muestra la tasa de reconocimiento por clase para las distintas variantes del sistema de clasificación entrenado, sobre las bases de datos DGS-FS_1 y DGS-FS_2. Aquí se han empleado la misma referencia de colores que en la Figura 5.12 y en color *magenta* se ha adicionado la tasa de reconocimiento por clase reportada por los creadores de la base de datos RWTH German Fingerspelling Database, tal como se describió en la Sec. 3.2.1 y empleada según lo reportado en [41]. Estos resultados fueron reproducidos a partir de la matriz de confusión reportada en http://www-i6.informatik.rwth-aachen.de/aslr/fingerspelling.php y se han agregado aquí a modo de referencia.

Como puede observarse, la tasa de reconocimiento de las distintas variantes resultó fuertemente dependiente de la clase, incluso en los resultados reportados por Dreuw y cols., atribuyendo este hecho a la gran similitud existente entre algunas clases [41]. A modo de ejemplo, las clases 1 y 29 fueron siempre mal clasificadas a partir de las características 'cOP', mientras que las clases 6, 20 y 24 muestran una tasa del orden del 80 %, de modo prácticamente independiente de las características y el clasificador empleado. Ha quedado fuera de los alcances de esta tesis la proposición de experimentos que permitan explicar este fenómeno.

En la Tabla 5.1 se muestra el porcentaje de $accuracy_{\mathrm{UB}}$ correspondiente a cada una de las estrategias exploradas sobre DGS-FS-source_labels. Mediante KNN se hace referencia al clasificador por KNN, considerando el K óptimo especificado entre paréntesis.

	DGS-FS_1		DGS-FS_2	
Conjunto de características	SVM	KNN	SVM	KNN
'cInc-v3'	8.67	6.36 (30)	11.64	9.08 (60)
'cDH'	60.74	54.36 (30)	62.21	54.88 (30)
'cOP'	64.46	64.45 (7)	64.46	64.45 (7)
'cDHOP'	64.81	56.77 (30)	**66.34**	57.77 (30)

Tabla 5.1: $accuracy_{\mathrm{UB}}$ para cada una de las variantes exploradas sobre DGS-FS-source_labels.

Tomando como referencia los resultados reportados por Dreuw y cols. [41] en 2006 sobre la base de datos RWTH German Fingerspelling Database, se observa que la tasa de desempeño global obtenida fue similar –66.34 % *versus* 64.3 %–. No obstante, Dreuw y cols. realizan la clasificación a nivel de *video* a partir de un HMM alimentado con una serie de descriptores de deformación de las manos a lo largo de la secuencia, mientras que en este trabajo la clasificación se realizó a nivel de *frame* mediante *aprendizaje por transferencia* y el entrenamiento de un clasificador SVM. En relación a este último resultado, con base en la metodología seguida es posible garantizar que el desempeño reportado en la Tabla 5.1 resulta independiente del señante.

5.4.4. Resultados sobre TReLSU-HS

En la Figura 5.14 se muestra la tasa de reconocimiento por clase para las distintas variantes del sistema de clasificación entrenado, sobre las bases de datos TReLSU-HS_1 y TReLSU-HS_2 modificadas según se comentó en la Sec. 5.4.1. Aquí se han empleado los mismos colores que en la Figura 5.12.

Como puede observarse, la tasa de reconocimiento es fuertemente dependiente de la clase. En particular, prácticamente la totalidad de las muestras de las clases 1, 6, 10, 14, 23, 24, 28, 39, 41 y 46 fueron mal clasificadas.

En la Figura 5.15 se muestra la tasa de reconocimiento *versus* la cantidad de muestras por clase. Sobre cada una de las curvas se ha anotado la clase correspondiente. Asimismo en la leyenda de cada una de las gráficas se presenta el coeficiente r de Pearson y su $p-value$ asociado. Puede observarse que existe un grado moderado de correlación entre la cantidad de muestras y la tasa de reconocimiento,

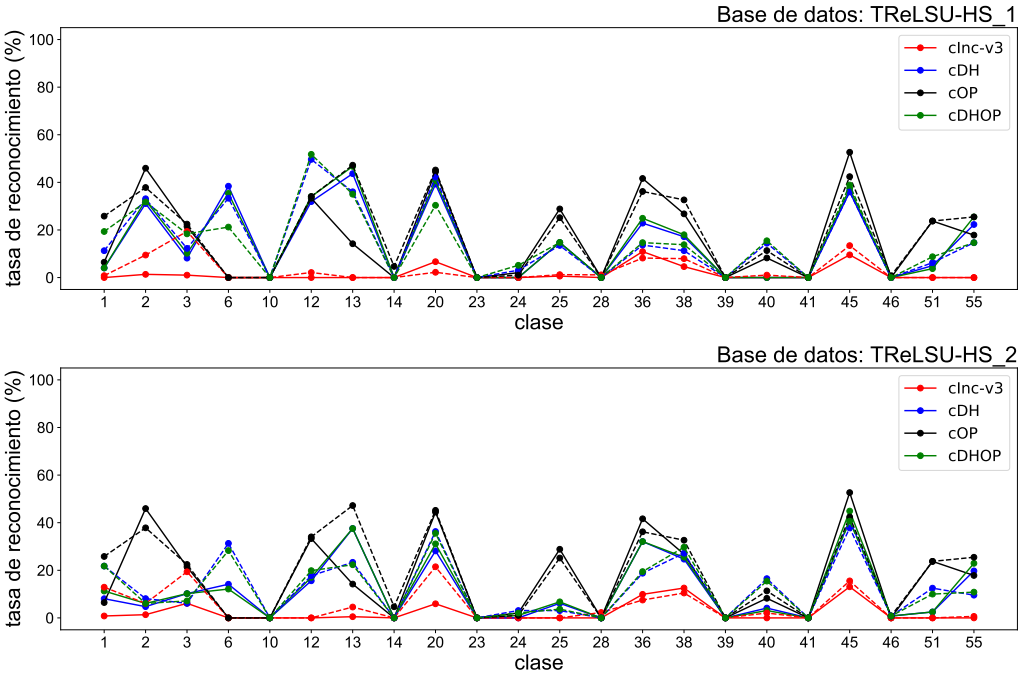


Figura 5.14: Tasa de reconocimiento por clase sobre TReLSU-HS. Referencia: SVM en línea sólida y clasificador por KNN en línea de trazos.

mostrando el clasificador por KNN valores de r menores que SVM, salvo para las características 'cOP'.

En la Tabla 5.2 se muestra el porcentaje de $accuracy_{\mathrm{UB}}$ correspondiente a cada una de las estrategias exploradas sobre TReLSU-HS. Mediante KNN se hace referencia al clasificador por KNN, considerando el K óptimo especificado entre paréntesis.

	TReLSU-HS_1		TReLSU-HS_2	
Conjunto de características	SVM	KNN	SVM	KNN
'cInc-v3'	1.58	3.04 (15)	2.28	4.67 (3)
'cDH'	14.30	15.04 (15)	11.34	12.45 (15)
'cOP'	16.71	**18.91** (7)	16.71	**18.91** (7)
'cDHOP'	14.90	15.18 (7)	12.04	12.44 (15)

Tabla 5.2: $accuracy_{\mathrm{UB}}$ para cada una de las variantes exploradas sobre TReLSU-HS.

A partir de los resultados presentados es posible concluir que las características y los clasificadores empleados mostraron un desempeño marcadamente inferior al correspondiente a DGS-FS, con un comportamiento fuertemente dependiente de la cantidad de muestras por clase, presentando mejores respuestas para aquellas clases con más muestras presentadas durante el entrenamiento. Este hecho es una consecuencia directa del desbalance de la base de datos TReLSU-HS y su consideración resulta sumamente importante en la proyección de este trabajo. La

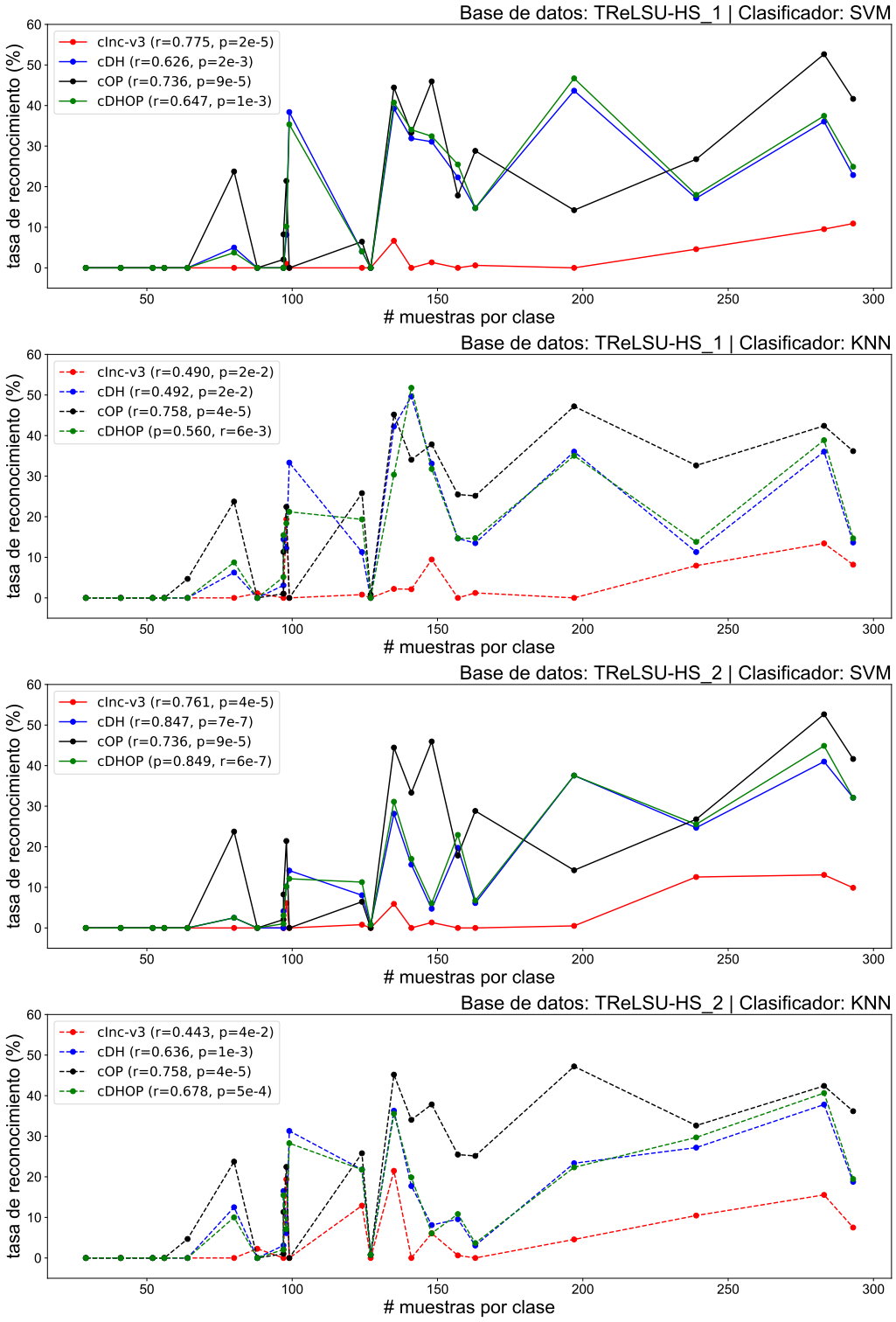


Figura 5.15: Tasa de reconocimiento *versus* cantidad de muestras por clase sobre TReLSU-HS. Referencia: SVM en línea sólida y clasificador por KNN en línea de trazos.

adquisición de una base de datos propia es de vital importancia para el desarrollo de un sistema de RALS uruguayo. En este sentido, se procurará la adquisición de una base de datos balanceada, teniendo especial cuidado en el muestreo uniforme de los datos sobre las distintas clases, sujetos y condiciones de registro.

Tanto en la base de datos DGS-FS como en TReLSU-HS, no se observaron diferencias estadísticamente significativas en el desempeño a partir de estas tres características. En particular, sobre DGS-FS se observó un desempeño levemente superior empleando SVM, salvo para 'cOP', cuyo desempeño fue muy similar para ambos clasificadores. Este hecho resulta muy interesante, teniendo en cuenta que a partir de 22 coordenadas 2D normalizadas –44 características– es posible obtener un desempeño muy similar al correspondiente a las 1024 características provistas por Deep Hand. A esto se le suma el hecho de que el desempeño del clasificador por KNN es comparable al clasificador SVM, siendo el primero de éstos un clasificador de una implementación mucho más sencilla. Este aspecto resulta muy atractivo y sugiere continuar la búsqueda de soluciones en este sentido.

5.5. Comentarios de fin de capítulo

A lo largo de este capítulo se presentaron las pruebas realizadas y los resultados obtenidos durante esta tesis de maestría.

En primer lugar, se verificó que el desempeño del sistema empleado para la realización de las pruebas sea similar al desempeño de la implementación en `TensorFlow` provista por Camgoz.

Luego, se realizó una serie de pruebas para evaluar el desempeño de Deep Hand frente a distintas bases de datos. En esta etapa fue necesario introducir una medida global de desempeño que considerara el desbalance de las clases en las bases de datos empleadas. De esta manera, se estimó una medida de exactitud más representativa del desempeño real de la etapa de clasificación. Bajo esta métrica, Deep Hand mostró un desempeño del orden del 30 % o inferior sobre las bases de datos consideradas, entre ellas la base de datos de prueba provista por sus autores. Durante esta etapa se propuso además un experimento para explorar la dependencia del desempeño de Deep Hand frente a cambios en la escala de la mano en la imagen de entrada, siendo prácticamente inapreciables las mejoras obtenidas.

En base a lo expuesto anteriormente, se decidió llevar a cabo el entrenamiento de un clasificador SVM y un clasificador por KNN bajo un esquema de aprendizaje por transferencia, empleando las características provistas por Deep Hand y OpenPose, de forma independiente y combinada sobre las bases de datos DGS-FS y TReLSU-HS.

Esta página ha sido intencionalmente dejada en blanco.

Capítulo 6

Conclusiones y perspectiva

Durante este proyecto fue posible la implementación de punta a punta de un sistema de reconocimiento de configuraciones manuales –propias de distintas lenguas de señas– y la evaluación del mismo bajo diferentes condiciones.

La realización de este trabajo permitió ganar experticia en diversas tareas inherentes al Reconocimiento Automático de la Lengua de Señas (RALS). En particular, se destacan el planteo del problema del RALS a distintos niveles de complejidad, el estudio de las alternativas existentes para su resolución y el uso de herramientas de desarrollo en el campo de la visión computacional.

En el marco de este proyecto se estudiaron las características fundamentales de las lenguas de señas, esto es, aspectos vinculados a la semántica de una seña como así también a la gramática de este tipo de lenguas. Ello permitió tomar noción de la complejidad propia de este medio de comunicación, y por tanto, de la complejidad ligada a su reconocimiento automático.

En términos generales, se observó que el RALS es frecuentemente abordado mediante una cadena de procesamiento, compuesta por las siguientes etapas: sensado, preprocesamiento, extracción de características y clasificación. Durante este trabajo se estudiaron distintas variantes para la implementación de cada una de estas etapas. En particular, se observa que las diferentes técnicas de sensado ofrecen distintos niveles de precisión en el RALS, al tiempo que acarrean una serie de restricciones para los señantes. En este sentido, es deseable que el señante no posea restricción alguna en cuanto a vestimenta, uso de guantes o condiciones del entorno de adquisición. De acuerdo con esto, se encontró que las soluciones basadas en aprendizaje profundo buscan independizar progresivamente el RALS de las restricciones mencionadas. Por esta razón, durante esta tesis de maestría se decidió optar y trabajar sobre este tipo de soluciones.

Mediante la revisión de las bases de datos y las métricas de desempeño, se conoció el material disponible para la implementación y evaluación de las distintas soluciones en este campo. Asimismo, esta búsqueda permitió tomar noción de los principales grupos activos en el RALS. Durante esta búsqueda no fue posible encontrar una base de datos de Lengua de Señas Uruguaya (LSU) para el reconocimiento automático. No obstante, las bases de datos estudiadas permitieron comprender los criterios y procedimientos seguidos para la adquisición de un

corpus con una aplicación particular. En virtud de ello, durante este trabajo se realizaron dos tareas. Por un lado, se conformó TReLSU-HS, una base de datos para el reconocimiento de configuraciones manuales propias de la LSU a partir de imágenes estáticas. Por otro lado, se sentaron las bases para la adquisición de una base de datos para el reconocimiento de LSU *a nivel de seña*, tomando un subconjunto de Léxico TReLSU como *corpus* de partida. En este sentido, será de vital importancia el trabajo interdisciplinario con expertos en el campo de la lingüística, a los fines de sistematizar correctamente el contenido a registrar y su etiquetado.

Durante la etapa de implementación en el marco de esta tesis de maestría se trabajó sobre la reproducción de un sistema de RALS para el reconocimiento de configuraciones manuales a partir de imágenes estáticas. En particular, el sistema *base* utilizado fue Deep Hand [75]. La mayor parte del tiempo asignado a esta etapa consistió en el planteo de la metodología para llevar a cabo la evaluación del sistema bajo distintas condiciones. La metodología seguida para la evaluación de Deep Hand implicó la selección y, eventualmente, la conformación de distintas bases de datos representativas del problema. Con el término 'conformación' se refiere a la obtención de las imágenes estáticas con su etiqueta correspondiente según las salidas de Deep Hand. En este sentido, fue necesario además proponer métodos de preprocesamiento de las imágenes que reunieran los requerimientos de Deep Hand, para lo cual se hizo uso de una librería abierta para la estimación de la postura corporal y manual denominada OpenPose. La reproducción del sistema Deep Hand fue posible y resultó consistente con los resultados reportados por sus autores. En términos generales, es posible concluir que el desempeño del sistema Deep Hand fue del orden del 30 % o inferior.

A los fines de mejorar el desempeño sobre las bases de datos empleadas, se propuso la realización de *aprendizaje por transferencia* a partir de las características extraídas por Deep Hand y OpenPose, de forma independiente y combinada, para el entrenamiento de un clasificador SVM y por KNN. Ello permitió obtener una tasa de reconocimiento global con *independencia de señante* máxima de 66.34 % sobre la base de datos DGS-FS –compuesta por 35 clases–, a partir de las características combinadas y empleando un clasificador SVM. Asimismo, este resultado fue puesto en perspectiva con los resultados reportados por Dreuw y cols. [41].

Sobre las distintas variantes de características y clasificadores evaluadas durante este trabajo, se destaca que el desempeño basado en las características provistas por OpenPose –vector de 42 dimensiones– mediante un clasificador por 7 vecinos más cercanos fue similar al mejor desempeño obtenido con la combinación de características de Deep Hand y OpenPose mediante un clasificador SVM –64.45 % *versus* 66.34 %, respectivamente–. En vista de la versatilidad de OpenPose y la simplicidad de un clasificador por KNN, este hecho sugiere continuar la búsqueda de soluciones en este sentido. A este hecho se le suma que una clasificación basada exclusivamente en los *keypoints* de la mano mediante un sistema como OpenPose podría prescindir de una etapa de segmentación previa.

En base a la revisión de la bibliografía es posible afirmar que el reconocimiento de configuraciones manuales actualmente cuenta con tasas de reconocimiento en el orden del 80 % [11] o superiores al 90 % [61,103,128], tanto a partir de un clasifica-

dor SVM como uno por KNN. No obstante, estos sistemas hacen uso de etapas de segmentación previas para la clasificación en un número reducido de clases [103], del cálculo de descriptores como SIFT y HOG [61, 128] sobre bases de datos pequeñas presegmentadas o con restricciones sobre la vestimenta de los señantes, o bien del reconocimiento basado en sub-unidades realizando una combinación de la información correspondiente a cada querema por medio de un HMM [11]. Ha quedado fuera de los alcances de esta tesis de maestría la reproducción de los trabajos citados en este párrafo y el planteo de experimentos que permitan una comparación válida contra los resultados obtenidos durante este trabajo.

Esta tesis de maestría constituye un punto de partida para la búsqueda de soluciones más específicas. Mediante la experiencia de este trabajo se comprendieron diversos detalles de las distintas soluciones existentes y propuestas. De manera general, es posible concluir que el RALS es un problema aún abierto, siendo los datos de naturaleza local una limitante importante en el campo. Se comentan a continuación algunas aristas sobre las cuales se continuará trabajando en pos de una solución efectiva al problema del RALS uruguayo.

En cuanto a las bases de datos locales, en primer lugar se estudiará la posibilidad del uso de técnicas de aumentado artificial de datos para el balanceo de clases de la base de datos TReLSU-HS existente. En segundo lugar, se abordará la adquisición de una base de datos propia, siguiendo los criterios expuestos en la Sec. 3.4 para el desarrollo de un sistema capaz de reconocer LSU a nivel de seña aislada.

Contando con una base de datos para el reconocimiento de LSU a nivel de seña, se deberán explorar las soluciones para el reconocimiento de secuencias, ya sea aquéllas empleadas desde el punto de vista clásico –HMM y DTW– como las soluciones basadas en *deep learning* –RNNs y variantes–. Asimismo, se evaluará la detección de las señas considerando la adición de los rasgos corporales y faciales. Una vez que se disponga de un sistema con un desempeño aceptable, se buscará la optimización de los tiempos y la adaptación del mismo en el desarrollo de *software* para la enseñanza primaria de la LSU, o bien como motor de búsqueda del diccionario en línea Léxico TReLSU, motivación inicial del presente trabajo.

Esta página ha sido intencionalmente dejada en blanco.

Apéndice A

Clases detectadas por Deep Hand

A continuación se exponen las clases detectadas por el sistema Deep Hand, reportada por Koller y cols. en [4]. Se trata de un conjunto de configuraciones manuales propias de las DGS. La etiqueta '0' representa la clase *basura* –o *garbage*–.

Clase	Configuración	Descripción	Clase	Configuración	Descripción
1		número ‘1’	31		letra ‘L’ *hook*
2		número ‘2’	32		dedo ‘medio’
3		número ‘3’	33		letra ‘M’
4		número ‘3’ *hook*	34		letra ‘N’
5		número ‘4’	35		letra ‘O’
6		número ‘5’	36		dedo ‘índice’
7		número ‘6’	37		dedo ‘índice’ *flex*
8		número ‘7’	38		dedo ‘índice’ *hook*
9		número ‘8’	39		palabra ‘*pincet*’
10		letra ‘A’	40		*ital*
11		letra ‘B’	41		*ital* más dedo ‘pulgar’
12		letra ‘B’ sin dedo ‘pulgar’	42		*ital* sin dedo ‘pulgar’
13		letra ‘B’ más dedo ‘pulgar’	43		*ital* abierto
14		letra ‘C’ *baby*	44		letra ‘R’
15		letra ‘O’ *baby*	45		letra ‘S’
16		palabra ‘*by*’	46		palabra ‘*write*’
17		letra ‘C’	47		palabra ‘*spoon*’
18		letra ‘D’	48		letra ‘T’
19		letra ‘E’	49		letra ‘V’
20		letra ‘F’	50		letra ‘V’ *flex*
21		letra ‘F’ abierta	51		letra ‘V’ *hook*
22		palabra ‘*fly*’	52		letra ‘V’ *hook* más dedo ‘pulgar’
23		palabra ‘*fly*’ sin dedo ‘pulgar’	53		letra ‘W’
24		letra ‘G’	54		letra ‘Y’
25		letra ‘H’	55		*umlaut* ‘AE’
26		letra ‘H’ *hook*	56		*umlaut* ‘AE’ más dedo pulgar
27		letra ‘H’ más dedo ‘pulgar’	57		palabra ‘*pincet*’ *double*
28		letra ‘I’	58		letra ‘O’ *baby double*
29		palabra ‘*jesus*’	59		letra ‘M’ (variante)
30		letra ‘K’	60		palabra ‘*jesus*’ más dedo ‘pulgar’

Tabla A.1: Clases de salida de Deep Hand. Reproducida de Tabla 2 de [4].

Apéndice B

Matrices de confusión de Deep Hand crudas

En este Anexo se incluyen las matrices de confusión *crudas* del sistema Deep Hand frente a las bases de datos segmentadas.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	**525**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2	0	0	**318**	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	**150**	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	6	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	**13**	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	3	0	0	0	0	0	**345**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0	0	0	0	0	**194**	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	2	0	0	0	0	0	0	0	0	0	0	28	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	0	0	0	1	0	0	0	0	20	0	**90**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	1	6	0	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**19**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
18	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**332**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	1	0	8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**19**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	4	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
31	0	0	15	1	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**13**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**393**	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	**0**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	**7**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**301**	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	2	0	0	0	0	0	0	10	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**103**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
49	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	**17**	0	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	3	**0**	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
53	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0
55	10	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**13**	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Tabla B.1: Matriz de confusión del sistema Deep Hand implementado en `TensorFlow` por Camgoz sobre la base de datos DH_Test.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	**525**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2	1	0	**318**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	**148**	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
4	1	0	1	5	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	**13**	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	2	0	0	0	0	0	**344**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3	0	0	0	0	0	0	0	0	0	0	**186**	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	3	0	0	0	0	0	0	0	0	0	0	27	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	4	0	0	0	0	0	0	0	0	0	0	16	0	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	1	8	0	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
18	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**328**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	1	0	7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**18**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
31	0	0	16	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**396**	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	**0**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	**3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	83	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**0**	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**299**	0	0	6	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	2	0	0	0	0	0	0	10	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**5**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**102**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
49	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	**17**	0	0	0	0	0	0	0	0	0	0	0
50	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	**0**	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
53	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	**0**	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0
55	8	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	**12**	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Tabla B.2: Matriz de confusión del sistema Deep Hand implementado en `PyTorch` sobre la base de datos DH_Test.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	54	**27**	0	0	0	0	15	0	0	0	0	10	0	12	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
2	33	2	**83**	1	0	0	10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0
3	55	0	6	**1**	0	0	30	0	0	0	0	0	0	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	11	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	10	0	0	0	0	0	**68**	0	0	0	0	3	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0
7	30	0	0	0	0	0	0	**0**	0	0	0	5	0	4	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	6	0	0	0	0	0	7	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	7	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	38	0	0	0	0	0	18	0	0	0	0	16	**46**	9	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	18	1	0	0	0	0	42	0	0	0	0	8	16	**91**	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
14	7	0	0	1	0	0	19	0	0	0	0	0	0	4	**5**	0	0	0	0	0	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	17	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	32	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	**62**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	18	2	0	0	0	0	1	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	2	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7	0	1	0	0	0	0
24	50	12	2	0	0	0	5	0	0	0	1	3	0	0	3	0	0	0	0	0	4	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	5	0	7	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	0
25	122	1	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0	0	**1**	0	0	0	1	0	0	0	0	0	0	6	0	1	0	2	0	0	0	0	0	0	0	0	20	0	2	0	0	0	1	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	44	4	1	0	0	0	0	0	0	0	1	1	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0	0	0	2	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	12	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	135	4	13	0	0	0	22	0	0	0	0	2	4	16	9	0	0	1	0	0	4	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	**67**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	75	22	2	0	0	0	19	0	0	0	0	0	4	17	20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	30	0	**5**	0	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	37	2	0	0	0	0
39	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	30	0	0	0	0	0	12	0	0	0	3	12	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**11**	0	16	8	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	0	0
41	29	0	1	2	0	0	19	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	9	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	165	30	4	0	0	0	5	0	0	0	8	4	0	19	0	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	2	0	7	1	0	**14**	6	0	0	0	0	0	0	0	0	10	0	0	0	0	0
46	62	4	0	0	0	0	7	0	0	0	5	6	0	1	1	0	0	3	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	**5**	0	0	0	0	0	0	0	0	5	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
49	28	0	0	2	0	0	1	0	0	0	0	0	0	0	4	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**5**	0	7	0	0	0	5	3	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
51	40	1	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	25	0	0	0	0	0
52	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
53	5	0	0	0	0	2	8	0	0	1	0	2	0	0	0	0	0	1	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	**0**	0	3	1	0	0	0	0
54	13	6	2	0	0	0	22	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0
55	105	3	1	0	0	0	13	0	0	0	0	1	2	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**27**	1	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Tabla B.3: Matriz de confusión del sistema Deep Hand implementado en PyTorch sobre la base de datos TReLSU-HS_1.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	73	**13**	1	0	0	0	9	0	0	0	15	0	0	12	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	50	0	**53**	0	0	0	20	0	0	0	0	0	0	11	1	0	0	0	0	0	0	0	5	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
3	86	0	3	**0**	0	0	4	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	11	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
6	15	0	1	1	0	0	**70**	0	0	0	0	0	0	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
7	24	0	0	0	0	0	0	**0**	0	0	0	16	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	26	0	0	0	0	0	1	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	59	0	0	0	0	0	22	0	0	0	0	12	**29**	10	0	0	0	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	35	0	0	0	0	0	26	0	0	0	0	2	13	**98**	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	5	0	0	0	0	0	23	0	0	0	0	0	0	6	**6**	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	16	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	35	0	0	0	0	0	33	0	0	0	0	0	1	0	0	0	0	0	0	0	**31**	0	0	0	0	0	0	0	0	11	0	0	0	0	0	7	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	23	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
24	70	0	3	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	**4**	0	0	0	0	6	0	0	0	0	0	0	0	0	4	0	1	0	2	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
25	119	0	0	1	0	0	3	0	0	0	0	0	1	6	0	1	0	0	0	0	5	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	4	0	0	0	1	0	0	4	0	0	0	0	0	11	0	0	0	0	1	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	51	2	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	2	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	144	1	8	0	0	0	12	0	0	0	0	5	2	26	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**80**	0	0	0	3	0	0	0	0	0	0	1	0	4	0	0	0	0	2	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	129	1	0	0	0	0	1	0	0	0	1	0	0	21	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35	0	**23**	0	0	0	0	0	0	0	0	0	0	1	0	0	0	20	0	0	0	0	0	0	0
39	24	0	0	0	0	0	1	0	0	0	0	1	0	5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	50	0	0	0	0	0	15	0	0	0	0	5	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**7**	0	3	13	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
41	28	0	0	0	0	0	19	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
42	5	0	0	0	0	0	1	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	248	3	0	0	0	4	0	0	0	0	0	1	0	15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	**2**	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
46	84	0	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	**3**	0	0	0	0	0	0	0	28	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	28	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	6	0	0	0	12	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
51	40	3	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	21	1	0	0	0	0	0	0
52	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
53	6	0	0	0	0	5	7	0	4	0	0	3	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	**1**	0	0	0	0	0	0	0	0
54	13	3	1	0	0	0	20	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0
55	106	1	2	0	0	0	15	0	0	0	0	0	0	3	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**27**	1	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0

Tabla B.4: Matriz de confusión del sistema Deep Hand implementado en `PyTorch` sobre la base de datos TReLSU-HS_2.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	66	2	**171**	18	0	0	14	0	0	0	0	0	0	6	1	0	0	0	0	0	0	0	34	4	1	0	0	0	0	32	0	2	0	0	0	0	6	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	5	13	0	0	0	0
3	34	0	10	**103**	0	0	37	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	9	0	0	0	0
4	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	73	0	0	0	0	**9**	52	0	0	0	0	2	0	0	0	0	0	2	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	34	0	0	0	0	0
6	52	0	1	0	0	0	**257**	0	0	0	0	0	0	2	0	0	0	2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	20	2	0	0	0	0
7	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	62	4	2	0	0	0	0	0	0	1	**65**	10	0	3	1	0	0	27	0	0	4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	7	0	0	0	0	14	1	21	0	0	0	0	8	0	0	0	2	0	0	3	159	7	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	29	0	0	0	0	0	3	0	0	0	1	6	**0**	1	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	18	0	3	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	31	3	0	0	0	0	3	0	0	0	0	0	0	1	6	0	0	**21**	0	0	8	0	0	0	0	0	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	0	0	0	0	0
18	68	1	33	0	0	0	2	0	0	5	0	0	0	1	3	0	0	6	**0**	0	9	0	3	1	0	0	0	0	1	1	0	3	0	0	0	0	31	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	13	12	0	0	0	0
19	32	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	26	0	**6**	4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	110	0	0	0	0	0
20	17	0	0	4	0	0	37	0	0	0	0	0	0	1	0	0	0	0	0	0	**110**	0	10	0	0	0	0	0	0	4	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	5	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	63	10	11	0	0	0	4	0	0	2	2	0	0	0	8	0	0	0	0	0	13	0	5	0	**8**	0	0	0	0	0	0	13	0	0	0	3	1	0	0	16	6	0	0	4	0	0	0	0	0	4	0	0	0	0	0	19	1	0	0	0	0
25	303	1	4	1	0	0	27	7	3	9	0	20	0	11	0	0	0	74	0	0	36	0	0	0	0	**68**	0	0	0	0	2	2	0	0	0	0	7	0	0	0	1	0	1	0	0	0	1	0	0	4	0	2	0	0	0	10	4	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	174	6	2	1	0	0	3	0	0	17	8	3	0	8	0	0	0	8	0	0	1	0	3	1	0	0	0	0	**3**	0	0	8	0	0	0	1	0	0	0	0	1	8	24	0	0	0	0	4	0	0	0	0	0	0	15	54	50	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	110	0	30	14	0	0	18	0	0	22	1	1	0	3	3	0	0	2	0	0	16	0	9	0	1	7	0	0	0	0	**0**	8	0	0	0	0	125	0	0	0	3	0	1	0	0	0	0	0	0	16	0	0	0	0	0	14	4	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	31	2	1	0	0	0	0	0	0	0	6	0	0	0	0	0	0	17	0	4	1	0	0	0	0	0	0	0	0	0	0	1	0	**0**	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	134	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	122	6	0	0	0	0	0	0	0	0	2	0	0	2	6	0	0	26	0	0	13	0	0	0	0	0	0	0	0	0	0	5	0	0	0	**76**	0	0	0	0	18	0	6	1	0	0	0	0	0	0	0	1	0	0	0	101	0	0	0	0	0
36	224	4	6	0	0	0	11	0	0	0	1	17	0	11	1	1	0	1	0	0	3	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	**33**	0	1	0	1	0	0	0	0	0	0	0	0	1	0	3	0	0	0	21	2	0	0	0	0
37	85	0	4	2	0	0	4	0	0	0	1	2	0	1	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	81	**0**	0	0	1	0	5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
38	57	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	9	0	**5**	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	73	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	88	1	16	1	0	0	5	0	0	7	0	0	0	17	0	0	0	9	0	0	5	0	0	0	0	1	0	0	0	0	0	6	0	0	0	0	35	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0
45	39	3	2	0	0	0	0	0	0	1	3	0	0	0	3	0	0	3	0	0	7	0	0	0	2	0	0	0	0	0	0	1	0	0	0	11	0	0	0	0	6	0	1	0	0	**0**	0	0	0	0	0	1	0	0	0	120	1	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
48	124	3	24	0	0	0	5	0	0	0	0	0	0	6	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	2	11	0	0	0	0	0	**1**	0	0	0	0	0	0	3	1	0	0	0	0
49	37	0	12	105	0	0	1	0	8	0	0	0	0	0	1	0	0	2	0	0	13	0	1	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	0	0	0	0	0	5	1	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
53	56	0	0	30	0	10	37	1	10	1	0	0	0	0	0	0	0	3	0	0	16	0	0	0	0	0	0	0	0	3	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	2	0	**1**	0	27	3	0	0	0	0
54	33	0	3	0	0	0	14	0	0	17	0	1	0	8	0	0	0	3	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	**23**	7	61	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Tabla B.5: Matriz de confusión del sistema Deep Hand implementado en PyTorch sobre la base de datos DGS-FS_1.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	66	4	**144**	19	0	0	20	0	0	0	0	0	0	12	7	0	0	0	0	0	0	0	46	4	0	0	0	0	0	41	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	11	0	0	0	0
3	41	0	4	**104**	0	0	39	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	6	0	0	0	0
4	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	55	0	0	0	0	**13**	51	0	0	0	2	4	0	0	0	0	0	4	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	1	0	0	0	0
6	74	1	1	7	0	2	**208**	0	0	0	0	0	0	3	0	0	0	1	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	4	0	0	0	0
7	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	56	1	5	0	0	0	1	0	0	3	**101**	4	0	8	0	0	0	21	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	7	0	0	0	0	0	0	12	0	0	0	0	3	0	0	0	1	0	0	1	163	11	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	34	0	0	0	0	0	1	0	0	0	0	19	**0**	6	0	0	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	11	0	3	0	0	0	0	1	0	0	0	0	0	0	0	12	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	29	0	0	0	0	0	6	0	0	0	0	1	0	2	8	0	0	**43**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78	7	0	0	0	0
18	79	2	29	0	0	0	4	0	1	10	0	0	0	2	1	0	0	3	**0**	0	11	0	3	0	0	0	0	0	0	1	0	4	0	0	0	0	19	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	7	14	0	0	0	0
19	29	0	1	0	0	0	0	0	0	0	6	1	0	0	0	0	0	22	0	**2**	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	125	0	0	0	0	0
20	14	0	0	5	0	0	40	0	0	2	0	0	0	0	0	0	0	0	0	0	**124**	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	6	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	122	0	16	2	0	0	13	0	0	1	0	2	0	0	0	0	0	0	0	0	6	0	0	0	**0**	0	0	0	0	1	0	3	0	0	0	1	0	0	0	0	3	0	0	2	0	4	0	0	0	0	0	0	0	0	1	16	0	0	0	0	0
25	312	0	4	4	0	0	49	2	3	10	0	20	0	21	1	0	0	50	0	0	32	0	0	0	0	**12**	0	0	1	0	0	5	0	0	0	0	1	0	0	0	2	0	4	0	0	0	2	0	0	8	0	7	0	0	0	22	26	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	146	0	1	0	0	0	11	0	0	26	7	10	0	15	0	0	0	4	0	0	3	0	2	0	0	0	0	0	**4**	0	0	2	0	0	0	1	0	0	0	0	0	4	25	0	0	0	0	4	0	0	0	0	0	0	5	61	72	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	184	0	47	8	0	0	41	0	0	22	1	0	0	3	5	0	0	0	0	0	15	0	0	1	0	1	0	0	0	1	**0**	10	0	0	0	0	26	0	0	0	0	1	4	0	0	0	0	0	0	18	0	0	0	0	0	10	10	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	18	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	15	0	4	6	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	133	0	2	0	0	0	1	0	0	0	1	0	0	10	0	0	0	69	0	0	3	0	0	0	0	0	0	0	0	0	0	2	0	0	0	**32**	0	0	0	0	18	0	4	0	0	1	0	1	0	0	0	0	0	0	0	108	0	0	0	0	0
36	208	5	6	0	0	0	13	0	0	0	0	16	0	31	2	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**20**	0	2	0	0	2	0	0	0	0	0	0	0	0	0	3	0	0	0	24	11	0	0	0	0
37	125	0	3	0	0	0	17	0	0	0	1	8	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	31	**0**	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0
38	31	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	9	0	0	0	0	5	0	**5**	5	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	98	1	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	86	2	7	0	0	0	7	0	0	4	0	3	0	29	0	0	0	7	0	0	4	0	0	0	0	2	0	0	1	0	0	5	0	0	0	0	16	0	0	0	0	0	0	0	**0**	0	0	1	0	0	0	0	0	0	0	7	17	0	0	0	0
45	34	3	7	0	0	0	0	0	0	1	2	0	0	0	1	0	0	4	0	1	14	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	3	0	0	**0**	0	0	0	0	0	1	0	0	0	128	3	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
48	104	5	34	1	0	0	9	0	0	1	0	1	0	5	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	12	0	0	0	0	0	0	0	0	0	1	18	0	0	0	0	0	**0**	0	0	0	0	0	0	0	7	0	0	0	0
49	53	0	8	84	0	0	3	0	11	1	0	0	0	1	1	0	0	2	0	0	16	0	2	0	0	0	0	0	0	5	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	0	1	0	0	0	6	1	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
53	41	0	0	31	0	15	35	0	4	1	0	1	0	2	1	0	0	1	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0	**1**	0	35	1	0	0	0	0
54	36	1	3	0	0	0	26	0	0	15	0	1	0	5	0	0	0	2	0	0	0	0	7	0	0	0	0	0	0	3	0	2	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	**16**	7	75	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Tabla B.6: Matriz de confusión del sistema Deep Hand implementado en PyTorch sobre la base de datos DGS-FS_2.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	232	0	**0**	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0
3	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	221	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	271	0	0	0	0	0	1	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	288	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
18	160	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	105	0	0	0	0	0
19	157	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0
20	240	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	489	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	256	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	455	0	0	1	0	0	28	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	89	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	207	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	224	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	41	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	225	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0
38	234	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0
39	239	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	267	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0
45	221	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	58	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
49	260	0	0	0	0	0	5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	7	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
53	306	0	0	0	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0
54	193	0	0	0	0	0	50	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	24	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Tabla B.7: Matriz de confusión del sistema Deep Hand implementado en `PyTorch` sobre la base de datos ASL-FS.

Apéndice C

Etiquetado de las bases de datos según las clases de Deep Hand

En este Anexo se incluyen las tablas empleadas para el etiquetado de las muestras sobre las bases de datos TReLSU-HS, DGS-FS y ASL-FS.

Léxico TReLSU	PhWMSHS	Representación gráfica
@+1-B	1	
@^1+C	2	
@+1+E	3	
@-1+A	5	
@+1+A	6	
@-1'C	7	
@R1-A	10	
@R1+A	11	
@-1+B	12	
@+1+B	13	
@^1'A	14	
@'1'B	17	
@'1'F	20	
@-1+I	23	
@^1^E	24	
@-1+D	25	
@-1-A	28	
@'1'H	35	
@-1+C	36	
@-1'D	38	
@^1^A	40	
@^1^C	42	
@-1-C	45	
@R1'A	46	
@-1+F	49	
@-1'A	51	
@-1+G	53	
@+1-A	54	
@'1'C	55	

Tabla C.1: Etiquetado de las muestras de TReLSU-HS según las salidas de Deep Hand.

DGS-FS	PhWMSHS	Comentarios
1	10	letra ‘A’ de la DGS.
2	12	letra ‘B’ de la DGS.
3	17	letra ‘C’ de la DGS.
4	18	letra ‘D’ de la DGS.
5	19	letra ‘E’ de la DGS.
6	20	letra ‘F’ de la DGS.
7	37	letra ‘G’ de la DGS.
8	25	letra ‘H’ de la DGS.
9	28	letra ‘I’ de la DGS.
10	28	letra ‘J’ de la DGS. Equivale a letra ‘I’ más trayectoria de ‘J’.
11	30	letra ‘K’ de la DGS.
12	2	letra ‘L’ de la DGS.
13	33	letra ‘M’ de la DGS.
14	0	letra ‘N’ de la DGS, se asignó etiqueta 0 –clase *basura*–.
15	35	letra ‘O’ de la DGS.
16	30	letra ‘P’ de la DGS. Equivale a letra ‘K’ rotada.
17	24	letra ‘Q’ de la DGS.
18	44	letra ‘R’ de la DGS.
19	45	letra ‘S’ de la DGS.
20	48	letra ‘T’ de la DGS.
21	25	letra ‘U’ de la DGS. Equivale a letra ‘H’ rotada.
22	49	letra ‘V’ de la DGS.
23	53	letra ‘W’ de la DGS.
24	38	letra ‘X’ de la DGS.
25	54	letra ‘Y’ de la DGS.
26	36	letra ‘Z’ de la DGS. Equivale a etiqueta 36 más trayectoria ‘Z’.
27	10	*umlaut* ‘Ä’ de la DGS. Equivale a letra ‘A’ más trayectoria ‘↕’.
28	35	*umlaut* ‘Ö’ de la DGS. Equivale a letra ‘O’ más trayectoria ‘↕’.
29	25	*umlaut* ‘Ü’ de la DGS. Equivale a letra ‘U’ más trayectoria ‘↕’.
30	6	letra ‘SCH’ de la DGS.
31	36	número ‘1’ de la DGS.
32	2	número ‘2’ de la DGS.
33	3	número ‘3’ de la DGS.
34	5	número ‘4’ de la DGS.
35	6	número ‘5’ de la DGS.

Tabla C.2: Equivalencia de etiquetas empleadas para la clasificación de las muestras de DGS-FS mediante Deep Hand.

ASL-FS	PhWMSHS	Comentarios
0	10	letra ‘A’ de la ASL.
1	12	letra ‘B’ de la ASL.
2	17	letra ‘C’ de la ASL.
3	18	letra ‘D’ de la ASL.
4	19	letra ‘E’ de la ASL.
5	20	letra ‘F’ de la ASL.
6	37	letra ‘G’ de la ASL.
7	25	letra ‘H’ de la ASL.
8	28	letra ‘I’ de la ASL.
10	30	letra ‘K’ de la ASL.
11	2	letra ‘L’ de la ASL.
12	33	letra ‘M’ de la ASL.
13	0	letra ‘N’ de la ASL, se asignó etiqueta 0 –clase *basura*–.
14	35	letra ‘O’ de la ASL.
15	30	letra ‘P’ de la ASL. Equivale a letra ‘K’ rotada.
16	39	letra ‘Q’ de la ASL.
17	44	letra ‘R’ de la ASL.
18	45	letra ‘S’ de la ASL.
19	0	letra ‘T’ de la ASL, se asignó etiqueta 0 –clase *basura*–.
20	25	letra ‘U’ de la ASL. Equivale a letra ‘H’ rotada.
21	49	letra ‘V’ de la ASL.
22	53	letra ‘W’ de la ASL.
23	38	letra ‘X’ de la ASL.
24	54	letra ‘Y’ de la ASL.

Tabla C.3: Equivalencia de etiquetas empleadas para la clasificación de las muestras de ASL-FS mediante Deep Hand.

Apéndice D

Matrices de confusión de Deep Hand con etiquetado de origen

En este Anexo se muestran las matrices de confusión del sistema Deep Hand frente a las bases de datos ASL-FS y DGS-FS con sus etiquetas de origen.

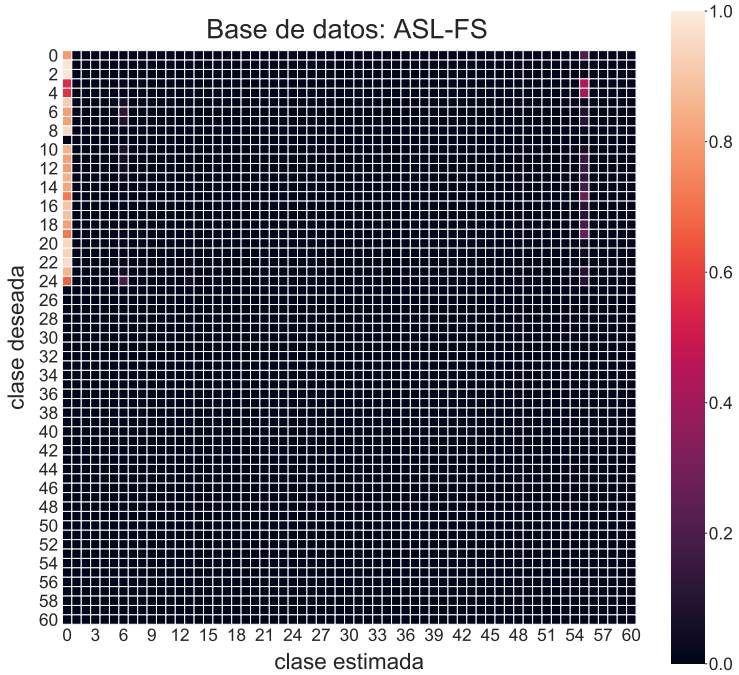


Figura D.1: Matrices de confusión del sistema Deep Hand con etiquetado de origen.

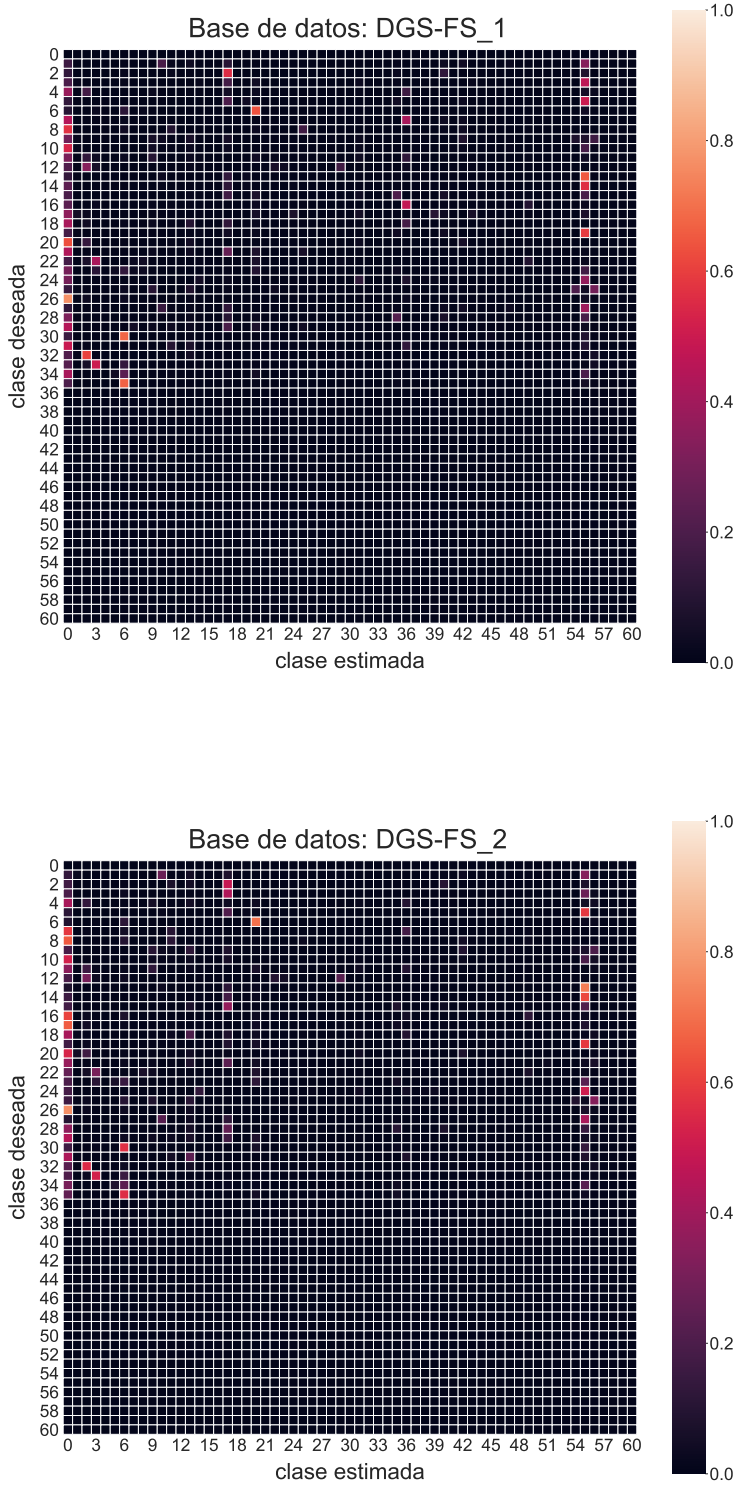


Figura D.2: Matrices de confusión del sistema Deep Hand con etiquetado de origen.

Apéndice E

Matrices de confusión de Deep Hand sin remoción de media por píxel

En este Anexo se muestran las matrices de confusión del sistema Deep Hand frente a las bases de datos segmentadas pero sin remoción de la media por píxel.

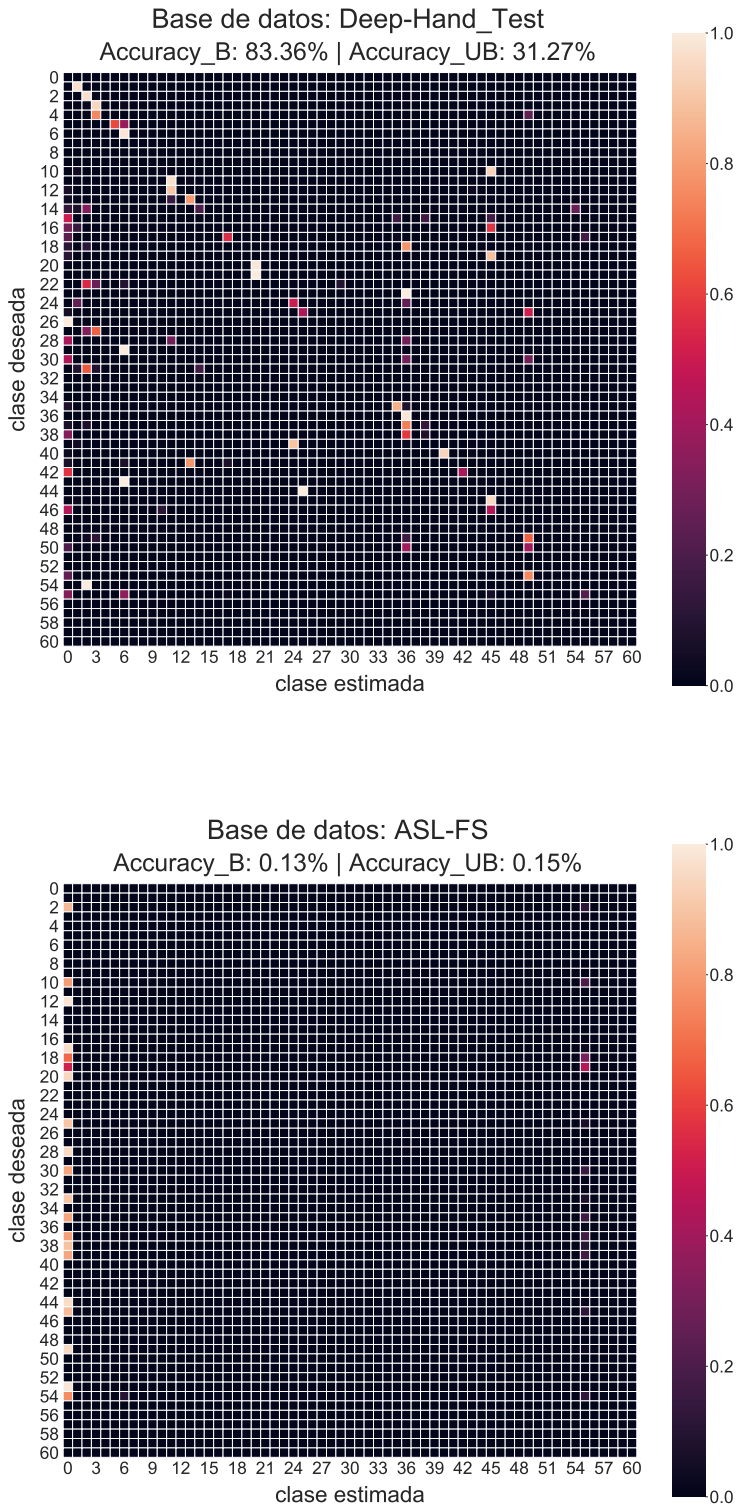
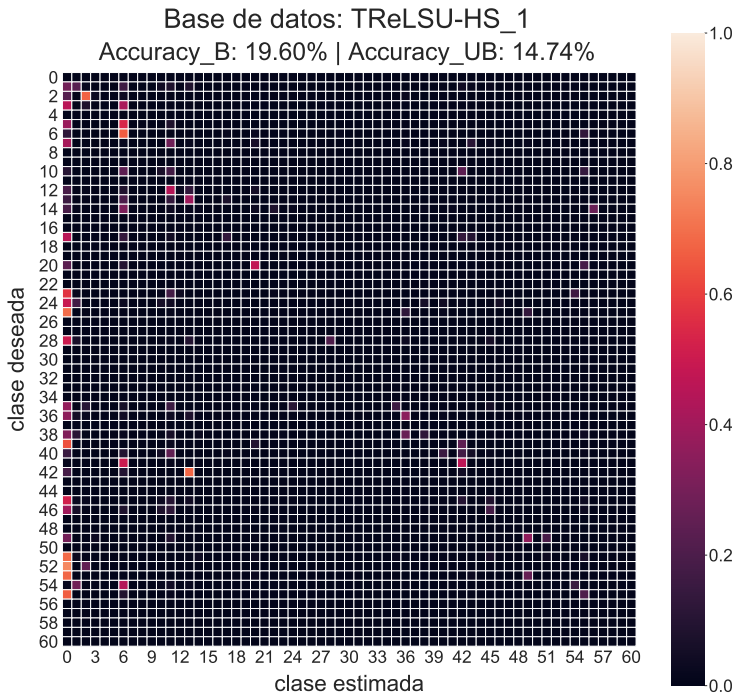


Figura E.1: Matrices de confusión del sistema Deep Hand sin remoción de media por píxel de las imágenes de entrada.



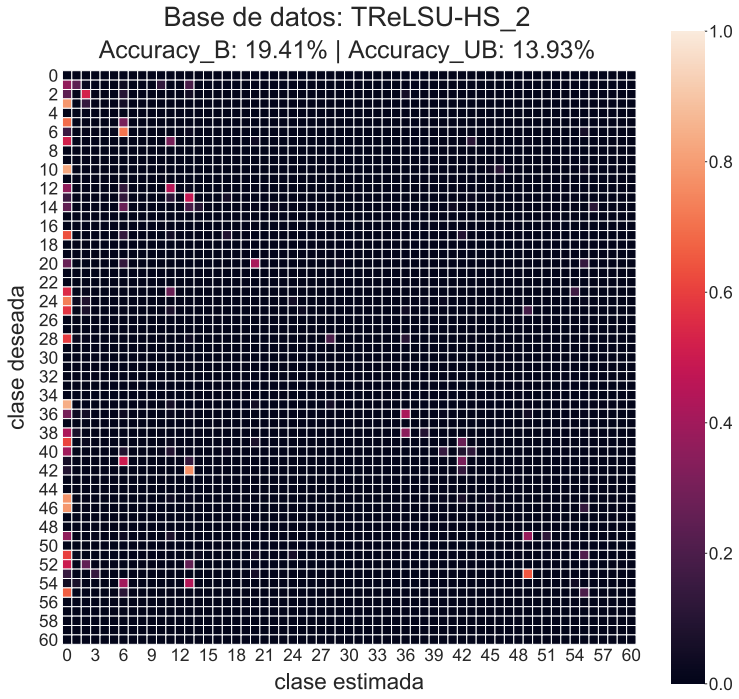


Figura E.2: Matrices de confusión del sistema Deep Hand sin remoción de media por píxel de las imágenes de entrada.

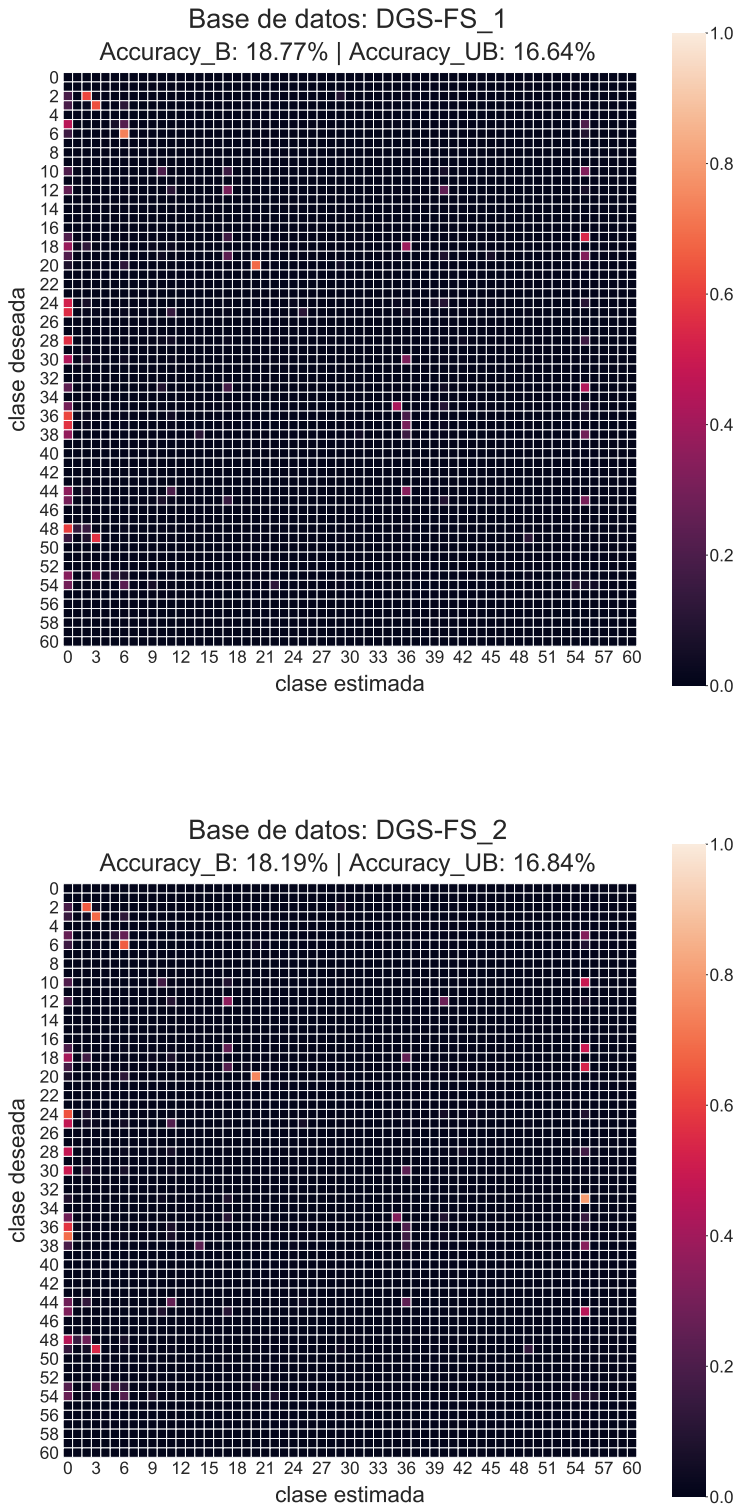


Figura E.3: Matrices de confusión del sistema Deep Hand sin remoción de media por píxel de las imágenes de entrada.

Apéndice F

Matrices de confusión de Deep Hand sobre ASL-FS con *zero-padding*

En este Anexo se muestran las matrices de confusión del sistema Deep Hand frente a la base de datos ASL-FS con distintos niveles p de *zero-padding*, según se explicó en la Sec. 5.2.

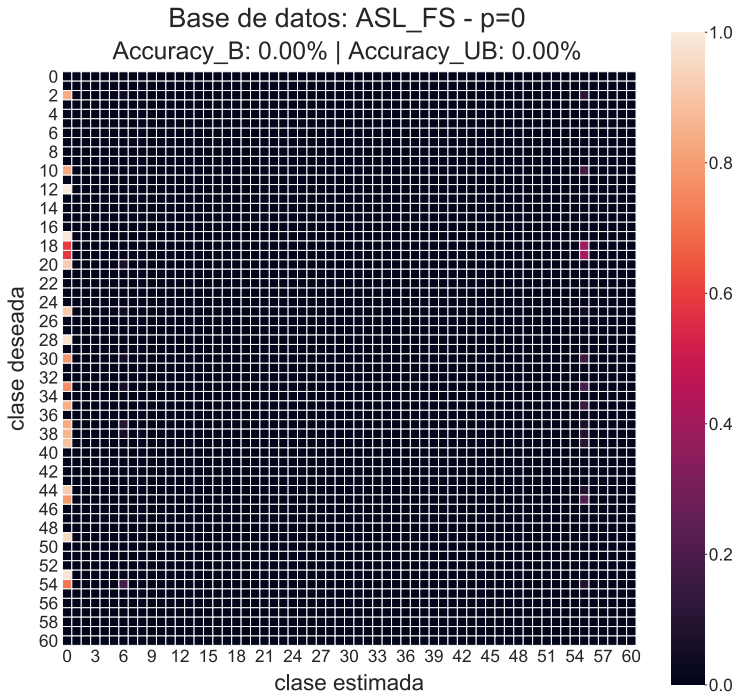


Figura F.1: Matriz de confusión del sistema Deep Hand sobre la base de datos ASL-FS, con $p = 0$.

Apéndice F. Matrices de confusión de Deep Hand sobre ASL-FS con *zero-padding*

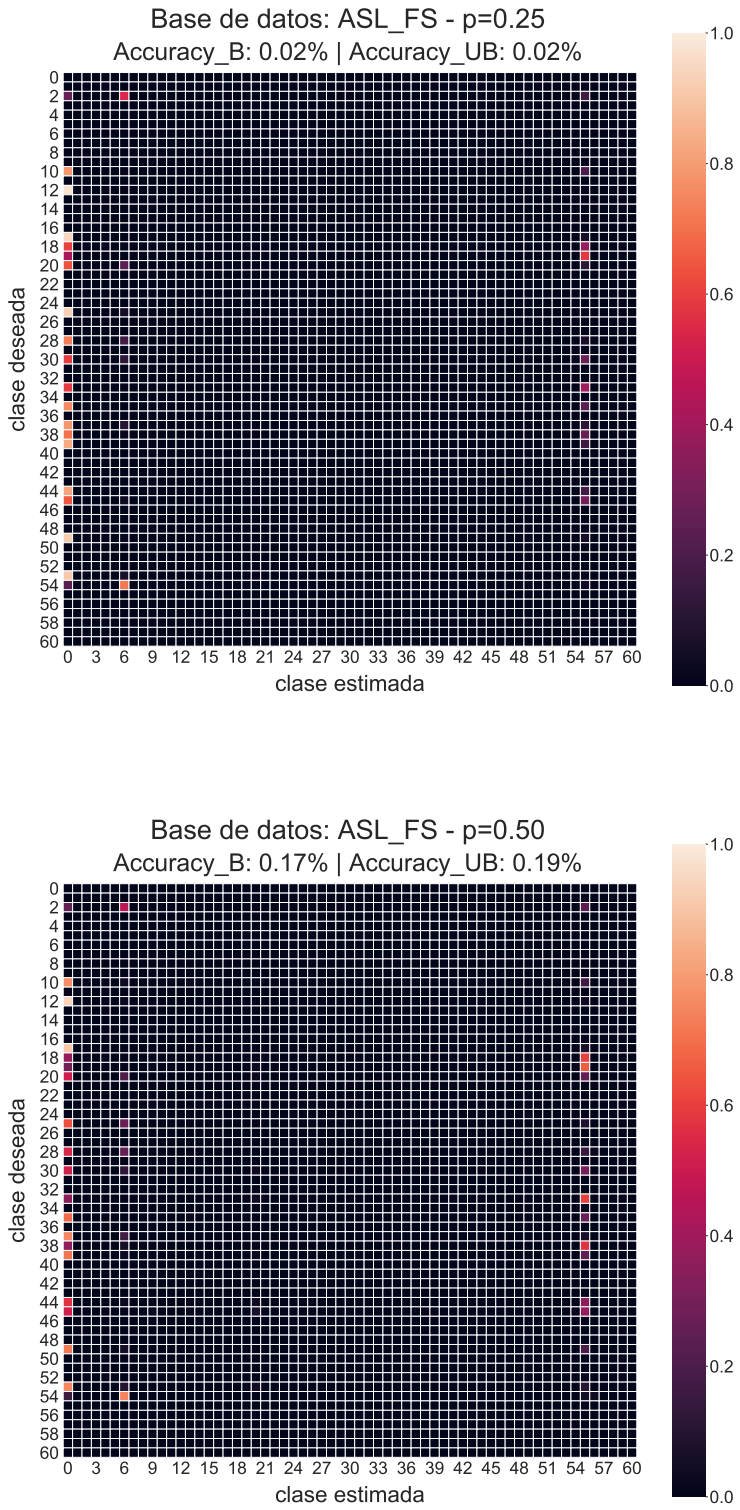


Figura F.2: Matriz de confusión del sistema Deep Hand sobre la base de datos ASL-FS, con $p = 0.25$ (arriba) y $p = 0.5$ (debajo).

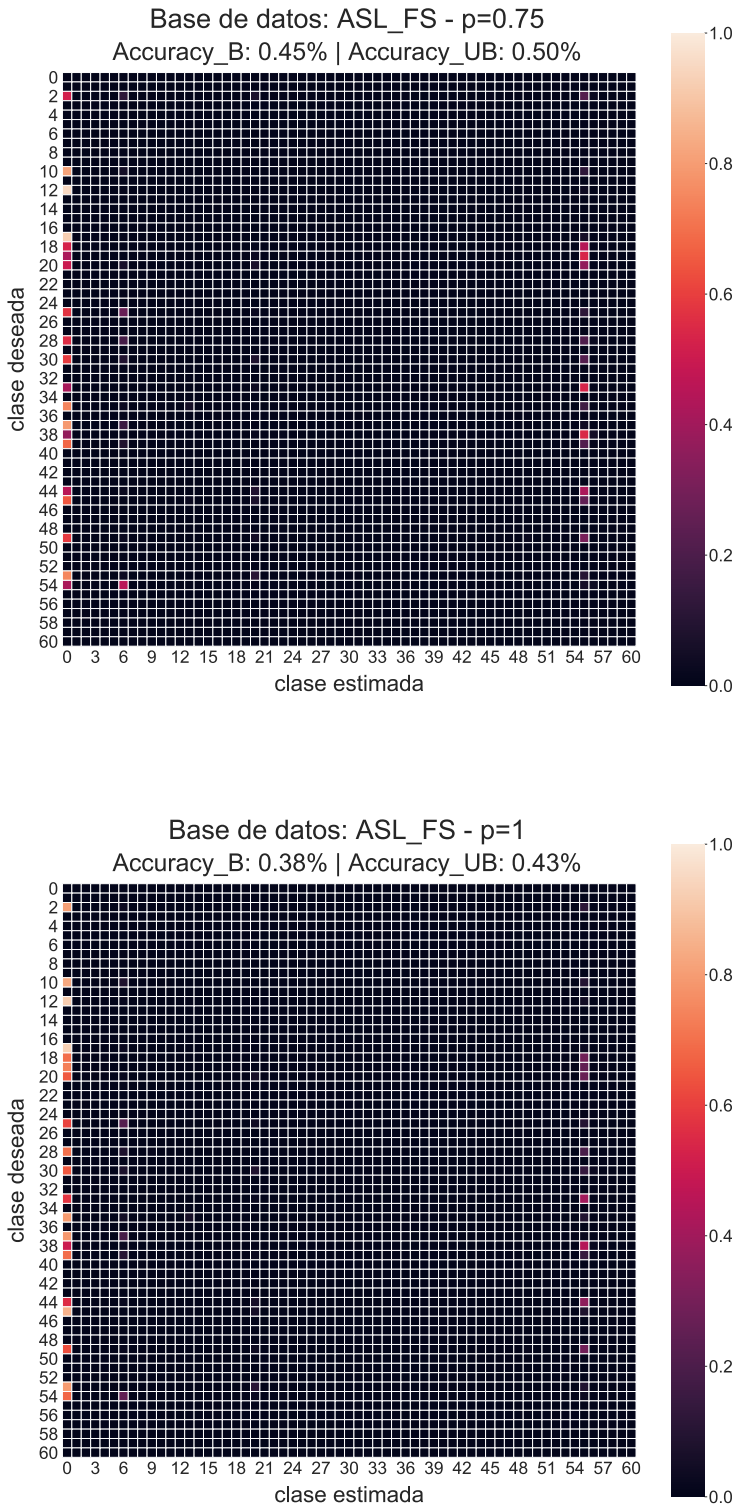


Figura F.3: Matriz de confusión del sistema Deep Hand sobre la base de datos ASL-FS, con $p = 0.75$ (arriba) y $p = 1.0$ (debajo).

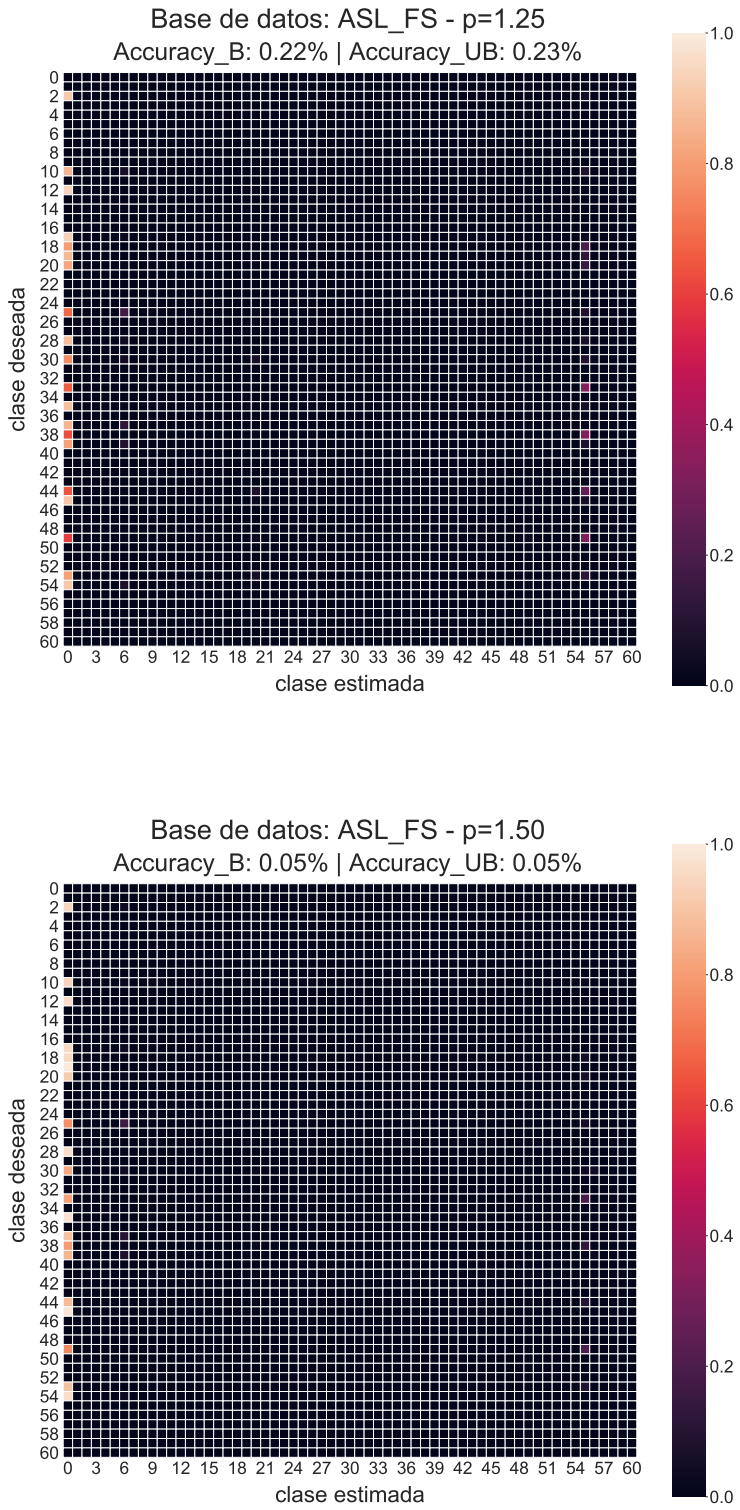


Figura F.4: Matriz de confusión del sistema Deep Hand sobre la base de datos ASL-FS, con $p = 1.25$ (arriba) y $p = 1.5$ (debajo).

Referencias

[1] CMU-Perceptual-Computing-Lab/openpose. Repositorio GitHub. https://github.com/CMU-Perceptual-Computing-Lab/openpose. Accedida: 2019-23-2.

[2] Leap Motion announces $50 million in Series C funding. https://haptic.al/leap-motion-announces-50-million-in-series-c-funding-a1a1f8c0440a. Accedida: 2018-12-31.

[3] Leap Motion, API Overview. https://developer-archive.leapmotion.com/documentation/csharp/devguide/Leap_Overview.html. Accedida: 2018-08-27.

[4] Public Hand Shape Data Set: RWTH-PHOENIX-Weather 2014 MS Handshapes. Oscar Koller, Human Language Technology & Pattern Recognition Group, RWTH Aachen University, Germany. https://www-i6.informatik.rwth-aachen.de/~koller/1miohands-data/. Accedida: 2019-2-25.

[5] Sign Language. Dr. Bill's ASL fingerspelling and handshape art. https://www.lifeprint.com/asl101/fingerspelling/abc-gifs/index.htm. Accedida: 2018-12-31.

[6] Sign Language Recognition Datasets. Repositorio GitHub de Facundo Quiroga. III-LIDI, Facultad de Informática, Universidad Nacional de La Plata. http://facundoq.github.io/unlp/sign_language_datasets/index.html. Accedida: 2019-2-7.

[7] Transfer Learning, CS231n Convolutional Neural Networks for Visual Recognition. Stanford University. http://cs231n.github.io/transfer-learning/. Accedida: 2019-03-05.

[8] World Health Organization, Deafness and hearing loss. http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accedida: 2018-09-09.

[9] Omar Al-Jarrah y Faruq A Al-Omari. Improving gesture recognition in the Arabic sign language using texture analysis. *Applied Artificial Intelligence*, 21(1):11–33, 2007.

[10] Omar Al-Jarrah y Alaa Halawani. Recognition of gestures in Arabic sign language using neuro-fuzzy systems. *Artificial Intelligence*, 133(1-2):117–138, 2001.

[11] Sílvia G. M. Almeida, Frederico G. Guimarães y Jaime A. Ramírez. Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors. *Expert Systems with Applications*, 41(16):7259–7271, 2014.

[12] Jörg Appenrodt, Ayoub Al-Hamadi, Mahmoud Elmezain y Bernd Michaelis. Data Gathering for Gesture Recognition Systems Based on Mono Color-, Stereo Color- and Thermal Cameras. En *International Conference on Future Generation Information Technology (FGIT)*, pp. 78–86. Springer, 2009.

[13] Oya Aran, Thomas Burger, Alice Caplier y Lale Akarun. A Belief-Based Sequential Fusion Approach for Fusing Manual and Non-Manual Signs. *Pattern Recognition*, 42(5):812–822, 2009.

[14] Ewout A. Arkenbout, Joost C. F. de Winter y Paul Breedveld. Robust Hand Motion Tracking through Data Fusion of 5DT Data Glove and Nimble VR Kinect Camera Measurements. *Sensors*, 15(12):31644–31671, 2015.

[15] Charlotte L. Baker-Shenk. A Microanalysis of the Nonmanual Components of Questions in American Sign Language. UC Berkeley Dissertations, Department of Linguistics, 1983.

[16] Igor L. O. Bastos, Michele F. Angelo y Angelo C. Loula. Recognition of Static Gestures applied to Brazilian Sign Language (Libras). En *28th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 305–312. IEEE, 2015.

[17] Herbert Bay, Tinne Tuytelaars y Luc Van Gool. SURF: Speeded Up Robust Features. En *European Conference on Computer Vision (ECCV)*, pp. 404–417. Springer, 2006.

[18] Sigal Berman y Helman Stern. Sensors for Gesture Recognition Systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3):277–290, 2012.

[19] Donald J. Berndt y James Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. En *KDD Workshop*, vol. 10, pp. 359–370. Seattle, WA, 1994.

[20] Stan Birchfield. *Image Processing and Analysis*. Cengage Learning, 2016.

[21] Cristopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, 2016.

[22] Matthew Brand. Coupled hidden Markov models for modeling interacting processes. The Media Lab, MIT, 1997.

[23] Helene Brashear, Thad Starner, Paul Lukowicz y Holger Junker. Using Multiple Sensors for Mobile Sign Language Recognition. Georgia Institute of Technology, 2003.

[24] Joseph D. Bronzino. *The Biomedical Engineering Handbook 1*. Electrical Engineering Handbook Series. Springer Berlin Heidelberg, 2000.

[25] Zhe Cao, Tomas Simon, Shih-En Wei y Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291-7299, 2017.

[26] Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg y Karen Emmorey. ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*, 49(2):784–801, 2017.

[27] Sait Celebi, Ali S. Aydin, Talha T. Temiz y Tarik Arici. Gesture Recognition Using Skeleton Data with Weighted Dynamic Time Warping. En *VISAPP (1)*, pp. 620–625, 2013.

[28] Feng-Sheng Chen, Chih-Ming Fu y Chung-Lin Huang. Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing*, 21(8):745–758, 2003.

[29] Justin K. Chen, Debabrata Sengupta y Rukmani R. Sundaram. Sign Language Gesture Recognition with Unsupervised Feature Learning. CS229: Machine Learning. Project Final Report. Stanford University, 2011.

[30] Zhi-hua Chen, Jung-Tae Kim, Jianning Liang, Jing Zhang y Yu-Bo Yuan. Real-Time Hand Gesture Recognition Using Finger Segmentation. *The Scientific World Journal*, pp. 1–9, 2014.

[31] Ming J. Cheok, Zaid Omar y Mohamed H. Jaward. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10(1):131–153, 2019.

[32] Ching-Hua Chuan, Eric Regina y Caroline Guardino. American Sign Language Recognition Using Leap Motion Sensor. En *13th International Conference on Machine Learning and Applications (ICMLA)*, pp. 541–544. IEEE, 2014.

[33] Necati C. Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney y Richard Bowden. Neural Sign Language Tsranslation. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7784–7793, 2018.

[34] Helen Cooper, Brian Holt y Richard Bowden. Sign Language Recognition. *Visual Analysis of Humans*, pp. 539–562. Springer, 2011.

[35] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault y Richard Bowden. Sign Language Recognition using Sub-Units. *Journal of Machine Learning Research*, 13:2205–2231, 2012.

[36] Andrea Corradini. Dynamic time warping for off-line recognition of a small gesture vocabulary. En *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 82–89. IEEE, 2001.

[37] D. McKee, R. McKee, S. P. Alexander y L. Pivac. The Online Dictionary of New Zealand Sign Language. `http://nzsl.vuw.ac.nz/`. Accedida: 2018-11-26.

[38] Nasser H. Dardas y Nicolas D. Georganas. Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. *IEEE Transactions on Instrumentation and Measurement*, 60(11):3592–3607, 2011.

[39] Trevor Darrell y Alex Pentland. Space-time gestures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 335–340. IEEE, 1993.

[40] Laura Dipietro, Angelo M. Sabatini y Paolo Dario. A Survey of Glove-Based Systems and Their Applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(4):461–482, 2008.

[41] Philippe Dreuw, Thomas Deselaers, Daniel Keysers y Hermann Ney. Modeling Image Variability in Appearance-Based Gesture Recognition. En *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing*, pp. 7–18, 2006.

[42] Philippe Dreuw, Jens Forster, Thomas Deselaers y Hermann Ney. Efficient approximations to model-based joint tracking and recognition of continuous sign language. En *8th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–6, 2008.

[43] Philippe Dreuw, Jens Forster y Hermann Ney. Tracking benchmark databases for video-based sign language recognition. En *European Conference on Computer Vision (ECCV)*, pp. 286–297. Springer, 2010.

[44] Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stan Sclaroff y Hermann Ney. Benchmark Databases for Video-Based Automatic Sign Language Recognition. En *Language Resources and Evaluation Conference (LREC)*, pp. 1115–1120, 2008.

[45] Philippe Dreuw y Hermann Ney. Towards Automatic Sign Language Annotation for the ELAN Tool. En *3rd Workshop on the Representation and Processing of Sign Languages*, pp. 50–53, 2012.

[46] Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi y Hermann Ney. Speech Recognition Techniques for a Sign Language Recognition System. En *8th Annual Conference of the International Speech Communication Association (ISCA)*, pp. 2513–2516, Antwerp, Belgium, August 2007. ISCA best student paper award Interspeech 2007.

[47] Philippe Dreuw, Daniel Stein, Thomas Deselaers, David Rybach, Morteza Zahedi, Jan Bungeroth y Hermann Ney. Spoken Language Processing Techniques for Sign Language Recognition and Translation. *Technology and Disability*, 20(2):121–133, 2008.

[48] Gareth J. Edwards, Timothy F. Cootes y Christopher J. Taylor. Face recognition using active appearance models. En European Conference on Computer Vision (ECCV), pp. 581–595. Springer, 1998.

[49] Paula Escudeiro, Nuno Escudeiro, Rosa Reis, Jorge Lopes, Marcelo Norberto, Ana Bela Baltasar, Maciel Barbosa y José Bidarra. Virtual Sign – A Real Time Bidirectional Translator of Portuguese Sign Language. *Procedia Computer Science*, 67:252–262, 2015.

[50] Gaolin Fang, Xiujuan Gao, Wen Gao y Yiqiang Chen. A novel approach to automatically extracting basic units from Chinese sign language. En *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 454–457, 2004.

[51] Li Fei-Fei y Pietro Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pages 524–531. IEEE, 2005.

[52] S. Sidney Fels y Geoffrey E. Hinton. Glove-Talk: a neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on Neural Networks*, 4(1):2–8, 1993.

[53] Pedro M. Ferreira, Jaime S. Cardoso y Ana Rebelo. On the role of multimodal learning in the recognition of sign language. *Multimedia Tools and Applications*, pp. 1–22. Springer, 2018.

[54] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater y Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. En *Language Resources and Evaluation Conference (LREC)*, pp. 3785–3789, 2012.

[55] Nancy Frishberg. Arbitrariness and Iconicity: Historical Change in American Sign Language. *Language*, 51(3):696–719, 1975.

[56] Srujana Gattupalli, Amir Ghaderi y Vassilis Athitsos. Evaluation of Deep Learning based Pose Estimation for Sign Language Recognition. En *9th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, paper no. 12, 2016.

[57] Binyam Gebrekidan Gebre, O. A. Crasborn, Peter Wittenburg, S. V. Drude y Tom Heskes. Unsupervised Feature Learning for Visual Sign Language Identification. En *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 370–376, 2014.

[58] Serkan Genç, Muhammet Başţan, Uğur Güdükbay, Volkan Atalay y Özgür Ulusoy. HandVR: a hand-gesture-based interface to a video retrieval system. *Signal, Image and Video Processing*, 9(7):1717–1726, 2015.

[59] Rafael C. Gonzalez y Richard E. Woods. *Digital Image Processing*. Prentice Hall, 2008.

[60] Ian Goodfellow, Yoshua Bengio y Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[61] Bhumika Gupta, Pushkar Shukla y Ankush Mittal. K-nearest correlated neighbor classification for Indian sign language gesture recognition using feature fusion. En *International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–5. IEEE, 2016.

[62] Simon Hadfield y Richard Bowden. Generalised Pose Estimation Using Depth. En *European Conference on Computer Vision (ECCV)*, pp. 312–325. Springer, 2010.

[63] Rudy Hartanto, Adhi Susanto y Paulus Insap Santosa. Real time static hand gesture recognition system prototype for Indonesian sign language. En *6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 1–6. IEEE, 2014.

[64] R. Hartley y A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[65] Alan Julian Izenman. *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer, 2008.

[66] Jette H. Kristoffersen, Thomas Troelsgård, Anne Skov Hardell, Bo Hardell, Janne Boye Niemelä, Jørgen Sandholt y Maja Toft. Ordbog over Dansk Tegnsprog. `http://www.tegnsprog.dk`. Accedida: 2018-11-26.

[67] Ulrich von Agris, Jörg Zieren, Ulrich Canzler, Britta Bauer y Karl-Friedrich Kraiss. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6(4), 323–362, 2008.

[68] Saba Joudaki, Dzulkifli bin Mohamad, Tanzila Saba, Amjad Rehman, Mznah Al-Rodhaan y Abdullah Al-Dhelaan. Vision-Based Sign Language Classification: A Directional Review. *IETE Technical Review*, 31(5):383–391, 2014.

[69] Daniel Kelly, John McDonald y Charles Markham. A person independent system for recognition of hand postures used in sign language. *Pattern Recognition Letters*, 31(11):1359–1368, 2010.

[70] Aisha U. Khan y Ali Borji. Analysis of Hand Segmentation in the Wild. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4710–4719. IEEE, 2018.

[71] Alireza Khotanzad y J.-H. Lu. Classification of invariant image representations using a neural network. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(6):1028–1038, 1990.

[72] Taehwan Kim. American Sign Language fingerspelling recognition from video: Methods for unrestricted recognition and signer-independence. PhD Thesis, Toyota Technological Institute at Chicago, Illinois, USA, 2016.

[73] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. En *Appears in the International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 14, no. 2, pp. 1137–1145. 1995.

[74] Oscar Koller, Hermann Ney y Richard Bowden. Deep Learning of Mouth Shapes for Sign Language. En *IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 85–91. IEEE, 2015.

[75] Oscar Koller, Hermann Ney y Richard Bowden. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3793–3802. IEEE, 2016.

[76] Dimitrios Konstantinidis, Kosmas Dimitropoulos y Petros Daras. Sign language recognition based on hand and body skeletal data. En *The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pp. 1–4. IEEE, 2018.

[77] P. Pramod Kumar, Prahlad Vadakkepat y Ai Poh Loh. Hand posture and face recognition using a fuzzy-rough approach. *International Journal of Humanoid Robotics*, 7(3):331–356, 2010.

[78] Pradeep Kumar, Himaanshu Gauba, Partha Pratim Roy y Debi Prosad Dogra. Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86:1–8, 2017.

[79] Ruslan Kurdyumov, Phillip Ho y Justin Ng. Sign Language Classification Using Webcam Images. CS229: Machine Learning. Project Final Report. Stanford University, 2011.

[80] Cheng Li y Kris M. Kitani. Pixel-Level Hand Detection in Ego-centric Videos. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3570–3577. IEEE, 2013.

[81] Yi Li. Hand gesture recognition using Kinect. MSc Thesis, University of Louisville, Kentucky, USA, 2012.

[82] Rung-Huei Liang y Ming Ouhyoung. A real-time continuous gesture recognition system for sign language. En *3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 558–567. IEEE, 1998.

[83] Jeroen F. Lichtenauer, Emile A. Hendriks y Marcel J. T. Reinders. Sign Language Recognition by Combining Statistical DTW and Independent Classification. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 30(11):2040–2046, 2008.

[84] Verónica López-Ludeña, Carlos González-Morcillo, Juan Carlos López, Roberto Barra-Chicote, Ricardo Córdoba y Rubén San-Segundo. Translating bus information into sign language for deaf people. *Engineering Applications of Artificial Intelligence*, 32:258–269, 2014.

[85] Sallandre Marie-Anne y Christian Cuxac. Iconicity in Sign Language: A Theoretical and Methodological Point of View. En *International Gesture Workshop: Gesture and Sign Language in Human-Computer Interaction*, pp. 173–180. Springer, 2001.

[86] Giulio Marin, Fabio Dominio y Pietro Zanuttigh. Hand gesture recognition with Leap Motion and Kinect devices. En *The IEEE International Conference on Image Processing (ICIP)*, pp. 1565–1569. IEEE, 2014.

[87] Giulio Marin, Fabio Dominio y Pietro Zanuttigh. Hand gesture recognition with jointly calibrated Leap Motion and depth sensor. *Multimedia Tools and Applications*, 75(22):14991–15015, 2016.

[88] Khoo Wei Ming y Surendra Ranganath. Representations for facial expressions. En *7th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, vol. 2, pp. 716–721. IEEE, 2002.

[89] Ross E. Mitchell, Travas A. Young, Bellamie Bachelda y Michael A. Karchmer. How Many People Use ASL in the United States?: Why Estimates Need Updating. *Sign Language Studies*, 6(3):306–335, 2006.

[90] Boris Mocialov, Graham Turner, Katrin Lohan y Helen Hastie. Towards Continuous Sign Language Recognition with Deep Learning. En *Workshop on the Creating Meaning With Robot Assistants: The Gap Left by Smart Devices*, 2017.

[91] Mohamed Mohandes, S. Aliyu y M. Deriche. Arabic Sign Language Recognition using the Leap Motion Controller. En *23rd International Symposium on Industrial Electronics (ISIE)*, pp. 960–965. IEEE, 2014.

[92] Carol Neidle, Augustine Opoku, Gregory Dimitriadis y Dimitris Metaxas. NEW Shared & Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank. En *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, Language Resources and Evaluation Conference (LREC), 2018.

[93] Carol Neidle, Ashwin Thangali y Stan Sclaroff. Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus.

En *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*. Language Resources and Evaluation Conference (LREC), 2012.

[94] Carol Neidle y Christian Vogler. A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface (DAI). En *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*. Language Resources and Evaluation Conference (LREC), 2012.

[95] Katsuhiko Ogata. *Modern Control Engineering*. Instrumentation and controls series. Prentice Hall, 2010.

[96] Sylvie C. W. Ong y Surendra Ranganath. Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 6:873–891, 2005.

[97] Gerardo Ortega. Iconicity and Sign Lexical Acquisition : A Review. *Frontiers in Psychology*, vol. 8, art. 1280, 2017.

[98] Vitor F. Pamplona, Leandro A. F. Fernandes, Joao Prauchner, Luciana P. Nedel y Manuel M. Oliveira. The Image-Based Data Glove. En *10th Symposium on Virtual and Augmented Reality (SVR)*, 2008.

[99] Murugesa Pandiyan Paulraj, Sazali Yaacob, Mohd Shuhanaz bin Zanar Azalan y Rajkumar Palaniappan. A phoneme based sign language recognition system using skin color segmentation. En *6th International Colloquium on Signal Processing & its Applications (CSPA)*, pp. 1–5. IEEE, 2010.

[100] Leonardo Peluso Crespi. Nueva versión del modelo de descripción fonológico TRELSU: matriz segmental-articulatoria, configuración y movimiento. *Lengua de Señas e Interpretación*, 5:63–95, 2014.

[101] David M. Perlmutter. What is sign language? Linguistic Society of America. `https://www.linguisticsociety.org/sites/default/files/Sign_Language.pdf`. Accedida: 2017-11-17.

[102] Pramod Kumar Pisharady y Martin Saerbeck. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141:152–165, 2015.

[103] Pramod Kumar Pisharady, Prahlad Vadakkepat y Ai Poh Loh. Attention Based Detection and Recognition of Hand Postures Against Complex Backgrounds. *International Journal of Computer Vision*, 101(3):403–419, 2013.

[104] Leigh Ellen Potter, Jake Araullo y Lewis Carter. The Leap Motion controller: a view on sign language. En *25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, pp. 175–178. ACM, 2013.

[105] Prashan Premaratne. *Human Computer Interaction Using Hand Gestures*. Cognitive Science and Technology. Springer Singapore, 2014.

[106] Nicolas Pugeault y Richard Bowden. Spelling it out: Real-time ASL fingerspelling recognition. En *IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 1114–1119. IEEE, 2011.

[107] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[108] Christopher Rajah. *Chereme-based recognition of isolated, dynamic gestures from South African sign language with Hidden Markov Models*. MSc Thesis, University of the Western Cape, South Africa, 2006.

[109] Ives Rey Otero y Mauricio Delbracio. Anatomy of the SIFT Method. *Image Processing On Line*, 4:370–396, 2014.

[110] Franco Ronchetti. *Reconocimiento de gestos dinámicos y su aplicación al lenguaje de señas*. Tesis de Doctorado, Universidad Nacional de La Plata, Argentina, 2016.

[111] Franco Ronchetti, Facundo Quiroga, César Armando Estrebou y Laura Cristina Lanzarini. Handshape recognition for Argentinian Sign Language using ProbSom. *Journal of Computer Science & Technology*, 16(1):1–5, 2016.

[112] Franco Ronchetti, Facundo Quiroga, César Armando Estrebou, Laura Cristina Lanzarini y Alejandro Rosete. LSA64: An Argentinian Sign Language Dataset. En *XXII Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016.

[113] Ethan Rublee, Vincent Rabaud, Kurt Konolige y Gary Bradski. ORB: an efficient alternative to SIFT or SURF. En *IEEE International Conference on Computer Vision (ICCV)*, pp. 2564–2571. IEEE, 2011.

[114] Rubén San-Segundo, R. Barra, R. Córdoba, L. F. D'Haro, F. Fernández, Javier Ferreiros, Juan Manuel Lucas, Javier Macías-Guarasa, Juan Manuel Montero y José Manuel Pardo. Speech to sign language translation system for Spanish. *Speech Communication*, 50(11-12):1009–1020, 2008.

[115] Daniel Scharstein and Richard Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[116] Marlon Sequeira, G. M. Naik, J. S. Parab y R. S. Gad. Sign language recognition using sEMG and IMU. En *10th Annual National Symposium on VLSI and Embedded Systems*, India, 2017.

[117] Amir Shahroudy, Jun Liu, Tian-Tsong Ng y Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019. IEEE, 2016.

[118] Khamar Basha Shaik, P. Ganesan, V. Kalist, B. S. Sathish y J. Merlin Mary Jenitha. Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space. *Procedia Computer Science*, 57:41–48, 2015.

[119] Tomas Simon, Hanbyul Joo, Iain Matthews y Yaser Sheikh. Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1145–1153. IEEE, 2017.

[120] Thad Starner y Alex Pentland. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. En *Motion-Based Recognition*, pp. 227–243. Springer, 1997.

[121] Thad Starner, Joshua Weaver y Alex Pentland. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[122] Thad Starner. Visual Recognition of American Sign Language Using Hidden Markov Models. MSc Thesis, Massachusetts Institute of Technology, Cambridge, USA, 1995.

[123] Ariel E. Stassi Danielli. Reconocimiento automático de *handshapes*. Proyecto Final de curso "Tratamiento de Imágenes por Computadora". Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Uruguay, 2018. `https://iie.fing.edu.uy/investigacion/grupos/gti/timag/trabajos/2018/lenguaje_senas/`. Accedida: 2019-7-4.

[124] William C. Stokoe Jr. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *Journal of Deaf Studies and Deaf Education*, 10(1):3–37, 2005.

[125] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke y Andrew Rabinovich. Going Deeper with Convolutions. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[126] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens y Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[127] Luís Tarrataca, André C Santos y João MP Cardoso. The current feasibility of gesture recognition for a smartphone using J2ME. En *ACM Symposium on Applied Computing*, pp. 1642–1649. ACM, 2009.

[128] Alaa Tharwat, Tarek Gaber, Aboul Ella Hassanien, M. K. Shahin y Basma Refaat. SIFT-Based Arabic Sign Language Recognition System. En *Afro-European Conference for Industrial Advancement*, pp. 359–370. Springer, 2015.

[129] Ghassem Tofighi, S. Amirhassan Monadjemi y Nasser Ghasem-Aghaee. Rapid hand posture recognition using Adaptive Histogram Template of Skin and hand edge contour. En *6th Iranian Conference on Machine Vision and Image Processing (MVIP)*, pp. 1–5. IEEE, 2010.

[130] Andrea Vedaldi y Brian Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. En *18th ACM International Conference on Multimedia*, pp. 1469–1472. ACM, 2010.

[131] M. Vermeerbergen, L. Leeson y O. A. Crasborn. *Simultaneity in Signed Languages: Form and function*. Current Issues in Linguistic Theory. John Benjamins Publishing Company, 2007.

[132] Christian Vogler y Dimitris Metaxas. Parallel hidden Markov models for American sign language recognition. En *International Conference on Computer Vision (ICCV)*, vol. 1, pp. 116–122. IEEE, 1999.

[133] Ulrich von Agris, Daniel Schneider, Jörg Zieren y Karl-Friedrich Kraiss. Rapid Signer Adaptation for Isolated Sign Language Recognition. En *24th IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[134] Ulrich von Agris, Moritz Knorr y Karl-Friedrich Kraiss. The Significance of Facial Features for Automatic Sign Language Recognition. En *8th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–6. IEEE, 2008.

[135] Ulrich von Agris y Karl-Friedrich Kraiss. Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. *Gesture in Human-Computer Interaction and Simulation*, 2007.

[136] Manjula B Waldron y Soowon Kim. Isolated ASL sign recognition system for deaf persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–271, 1995.

[137] Chunli Wang, Wen Gao y Shiguang Shan. An approach based on phonemes to large vocabulary Chinese sign language recognition. En *5th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 411–416. IEEE, 2002.

[138] Robert Y. Wang y Jovan Popović. Real-Time Hand-Tracking with a Color Glove. *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3, art. 63, 2009.

[139] Shih-En Wei, Varun Ramakrishna, Takeo Kanade y Yaser Sheikh. Convolutional Pose Machines. En *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4732. IEEE, 2016.

[140] Frank Weichert, Daniel Bachmann, Bartholomäus Rudak y Denis Fisseler. Analysis of the Accuracy and Robustness of the Leap Motion Controller. *Sensors*, 13(5):6380–6393, 2013.

[141] David A. Winter. *Biomechanics and Motor Control of Human Movement*. John Wiley & Sons, 2009.

[142] Yaser Yacoob y Larry S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

[143] Hee-Deok Yang, Stan Sclaroff y Seong-Whan Lee. Sign Language Spotting with a Threshold Model Based on Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1264–1277, 2009.

[144] Jason Yosinski, Jeff Clune, Yoshua Bengio y Hod Lipson. How transferable are features in deep neural networks? En *Advances in Neural Information Processing Systems (NIPS)*, pp. 3320–3328. Curran Associates Inc., 2014.

[145] Yu Yuan y Kenneth Barner. An Active Shape Model Based Tactile Hand Shape Recognition with Support Vector Machines. En *40th Annual Conference on Information Sciences and Systems*, pp. 1611–1616. IEEE, 2006.

[146] Morteza Zahedi, Daniel Keysers, Thomas Deselaers y Hermann Ney. Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition. En *Joint Pattern Recognition Symposium*, pp. 401–408. Springer, 2005.

[147] Xu Zhang, Xiang Chen, Yun Li, Vuokko Lantz, Kongqiao Wang y Jihai Yang. A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 41(6):1064–1076, 2011.

[148] Zhengyou Zhang. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia*, 19(2):4–10, 2012.

[149] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35(4):399–458, 2003.

Esta página ha sido intencionalmente dejada en blanco.

Índice de tablas

Esta página ha sido intencionalmente dejada en blanco.

Índice de figuras

Esta es la última página.
Compilado el jueves 4 julio, 2019.
http://iie.fing.edu.uy/